A Revolution in Data Analysis: How new, very powerful, easy to use, Graphical Data Analysis tools and techniques can empower even novice Subject Matter Specialists

David L. DesJardins, U.S. Bureau of the Census, Washington, D.C.

ABSTRACT:

This paper outlines a framework for a revolution in our data analysis capability. This is because of new powerful computer capabilities married to equally powerful/flexible graphics software packages. To illustrate this new capability, it highlights one of these new graphical techniques for data analysis. A basic factor in this technique is the U.S. Census Bureau's new, easy to learn/understand, point-and-click graphical data analysis capability. This paper also suggests that current techniques for imputing missing or erroneous data values in the Bureau's surveys of Business and Industries can be markedly improved. Proposed is the use of Leverage Plots -- a technique now over a decade old -- but, because of three key developments in technology/capability at the Bureau today, a very applicable methodology to do this task.

BACKGROUND:

At the U.S. Bureau of the Census, I am expanding the everyday use of very powerful Exploratory Data Analysis (EDA) methods by teaching novice-level classes to our Survey Analysts. Three key factors make the introduction of these methods at this time a momentous opportunity:

* New, very powerful graphics software (such as JMP ® and INSIGHT ® from SAS Institute Inc.) now give our Subject Matter Specialists the easy to use (point-and-click) tools required to generate EDA graphs. More significantly, these graphs give our Analysts unique new insights into their data and the ability to interact in real time with the data that is displayed therein. In addition, Analysts no longer must learn the intricacies of programming or wait for systems development efforts to produce the custom software that they need to do this.

* Likewise, the current generation of relatively inexpensive, yet very powerful computer hardware (i.e., Pentiums and Unix workstations) gives our Analysts the ability to generate dozens of these very insightful graphs in a matter of mere minutes – a feat that would have taken months to do just a few years ago.

* By using the above hardware and software tools, we have developed new techniques that greatly enhance the speed and efficiency of our data analysis tasks. Accordingly, our Analysts no longer need to waste their time (and valuable subject matter expertise) trying to analyze their data with "blind" CPU-formatted algorithms and cumbersome, boring, tabular printouts. In addition, since graphs have the extraordinary ability to communicate across a wide area of expertise -- they can also often make even very sophisticated statistical concepts clear to laymen. As such, our Statisticians can now more quickly and effectively explain to our Analysts the fundamental concepts behind these new graphical data analysis techniques. The end result is that these data analysis/editing tasks can now be placed into the hands of the individuals who are best qualified to do them – the Subject Matter Specialists. And so, our Statisticians can now focus the majority of their time and efforts on improving our methodology.

This paper discusses only one of the many graphical data analysis techniques being used by the Author to introduce EDA to the Bureau. These graphical techniques have not only made our data editing/analysis tasks much easier and faster, but they also have given us special insights into our data. These new techniques are part of a much broader plan devised by the Author to ensure implementation of EDA methodology at the Bureau. This plan includes week-long EDA classes, easy to use EDA "cookbooks", seminars focused on innovations in graphic display and analysis, and a very active Graphics Users Group.

INTRODUCTION:

"The value of any statistical methodology is directly related to how easy it is to understand -- and inversely to how hard it is to implement."

ANON

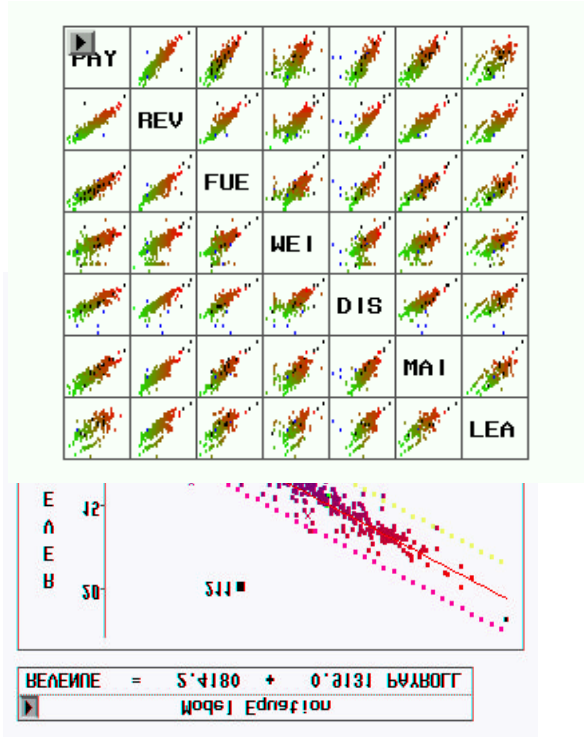
In the EDA class I currently teach at the Census Bureau, I have found that if my students do not clearly understand a technique, they often do not use it. In addition, I have found that graphically representing their data gives our Analysts unique insights into their current data analysis methodology -- often suggesting here-to-fore unforseen approaches to the data analysis tasks. Finally, I have found that our Subject Matter Specialists (who often have only a low/average level of statistical sophistication) are empowered by the insights gained by the graphic representation of their data.

Accordingly, one of the key things I do in this class is to use the power of graphs (as a universal communication medium) to help our Analysts understand often complex Statistical techniques. For instance, the EDA technique of "brushing" has proven to be a key factor in helping our Analysts see how the points in a number of different kinds of graphs interrelate. (Brushing highlights a subset of points in one graph -- which results in these same points simultaneously being highlighted in all the adjoining graphs.) By starting with a simple graph/relationship like a scatterplot, with brushing, students can often see and understand multi-variate interrelationships that are very complex. Please note how I use brushing in this paper -- as a supplement to the use of Leverage plots -- to help make this methodology more easily understood and implemented.

CURRENT METHODOLOGY:

Often, today at the Bureau, only a single variable is used to impute missing/erroneous company data values on our survey response forms. This single variable may be last year's reported values, or a closely related variable reported by the company on the current survey form. Typically, for editing purposes, a ratio of the differences between past year/current year responses are calculated (this is often referred to as "Ratio Editing"). Then, depending on the survey, something like the largest 100 of these differences are referred to the Analyst on a computer printout for follow-up. A standard formula is then used to impute the values of those industries that fail this edit, or to impute the missing values on the survey response forms.

For purposes of illustration, let's now take a set of artificial raw

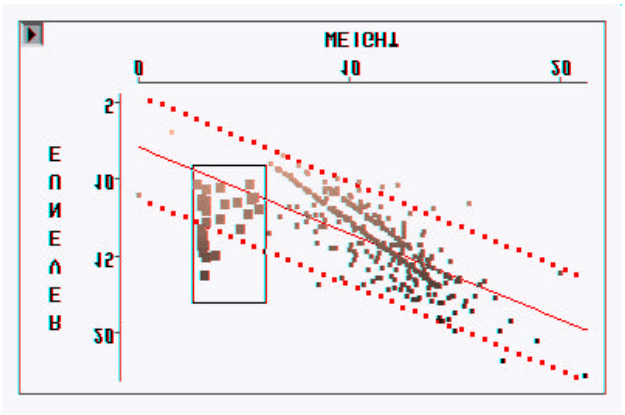


data values from the Bureau's Trucking Survey[1]. As is seen below, using a technique like Ratio Editing and a single comparison variable is often quite sufficient -- there is an excellent correlation between reported Revenues and Payroll for these companies. After eliminating the two obvious outlier points, perhaps we should use the indicated fit equation at the top of the graph and just stop here! (Points # 211 and # 150 are obvious outliers and have been flagged for follow up -- again, these are raw data values.)

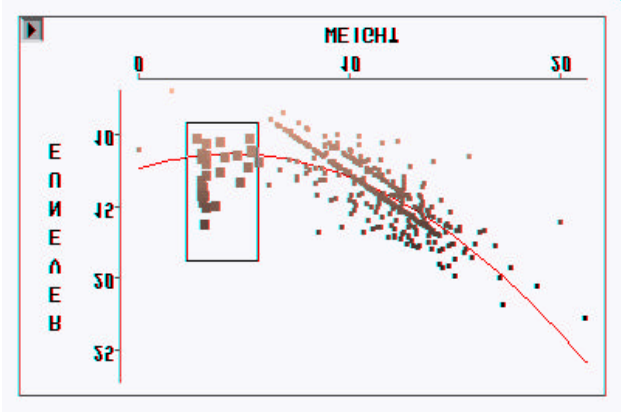
Our task here is to explore the use of the many additional variables from this survey form -- and to see if this might improve the accuracy of the imputed values. As can be seen from the scatterplot matrix of this data set (below), there is also a rather good correlation between reported Revenue and a number of the other variables that were reported on this Transportation Industry survey form: Fuel Cost, Weight of Goods, Distance Shipped, Truck Maintenance & Lease Costs.

But, why use more than one variable? Unfortunately, some of the data points in many survey sets are not so well behaved as our first example. Sometimes, whole subsets of points simply do not fit -- for instance, a hidden error exists or an unknown subset of the data (for instance, type of trucking) needs to be treated differently. In the example shown below, the highlighted points (with unusually high Revenue and low Shipping Weight) are simply out of place. The top graph shows a linear fit line with appropriate editing limits. Even with a non-linear fit (shown in the second graph), the eye can easily detect that there is a problem. (This is why we need to always look at our data -- a computer algorithm is blind! Even if we were to use a very sophisticated computer algorithm to edit these data, it is obvious that we could easily have eliminated good data values.) In this instance, further investigation disclosed that these highlighted companies were reporting Shipping Weight in tons instead of the requested format, pounds. (This key type of error had remained undetected until we discovered it with these graphic techniques.)

The erroneous points shown below could easily have been within the single variable data chosen for imputation (for instance, only the Weight values would be our single imputation variable for imputing missing/erroneous Revenue values). Accordingly, we are challenged to question the accuracy/capability of our current methodology. Even our other more sophisticated techniques, like Greenberg's SPEER or Multiple Imputation, would have a lot of problems with this kind of misreporting error -- garbage in, garbage out.



[1] *To protect confidentiality, this artificial, but representative, data set is used. The scale values of all data points of these graphs are in logs.*



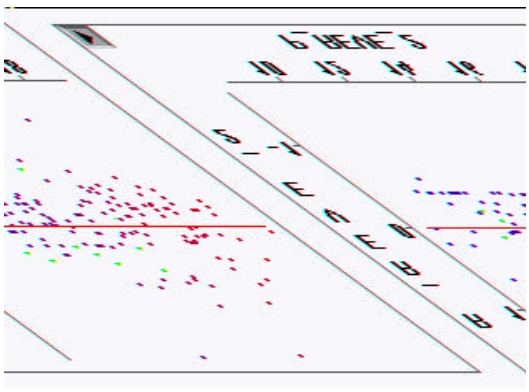
LEVERAGE PLOTS:
As shown above, because the erroneous tons/pounds points, if we were to use the Revenue/Weight fit line equation to impute missing values (without editing these misreported points), we would have introduced a significant error into our calculations. At this point, then, we need to ask if there is an easy to understand/implement methodology that will minimize the probability of this kind of an error. Required is a methodology that will quickly identify additional variables that could be used to impute missing/erroneous values on the survey form -- AND, a methodology that allows us to easily spot points (companies) that have an undue influence on the fit of all the other points!

I propose to use a methodology that utilizes as many of the related variables to impute missing values as possible. The scatterplot matrix (above) showed us that a number of these variables have a rather good correlation with one another. Below is the fit equation and key results from the statistical summary from a general fit between the reported Revenues and all of these variables. Aside from the Weight statistic, these "blind" statistics seem pretty good. Even the residual plot of this multivariate fit (shown below) is rather well behaved -- it has only a few outlier points that seem to need editing.

Revenue = 2.088 + 0.1042 Fuelcst + 0.5610 Payroll + 0.0279 Weight + 0.0491 Distance + 0.1016 Maintenance + 0.1485 Leasecst

VARIABLE	Prob > F	T Statistic
Fuelcost	0.0081	2.6797
Payroll	0.0001	16.1284
Weight	0.0804	1.7589
Distance	0.0265	2.2391
Maintenance	0.0020	3.1434
Leasecost	0.0001	10.4743

This is a plot of the fit residuals for all of these variables:

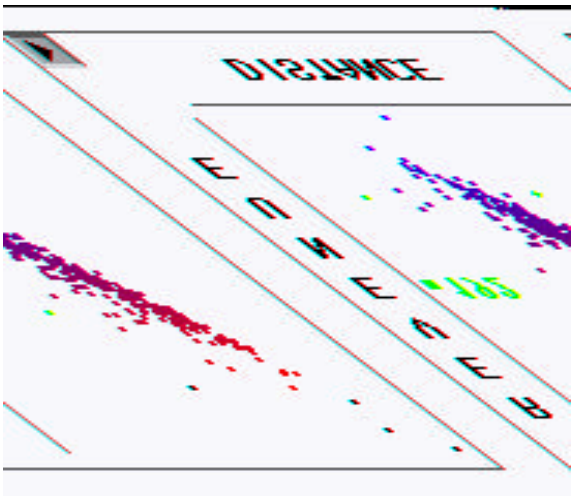


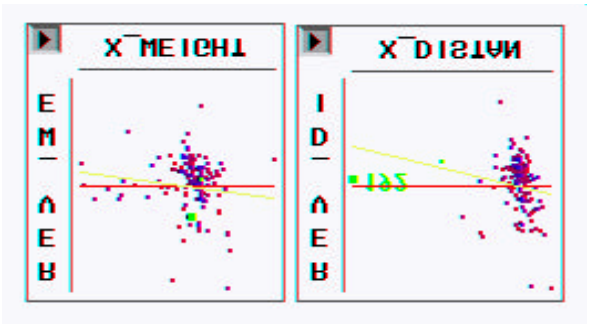
I propose the use of Sall's updated Leverage plots to let Analysts supplement our current methodology to improve the accuracy of our imputed values.

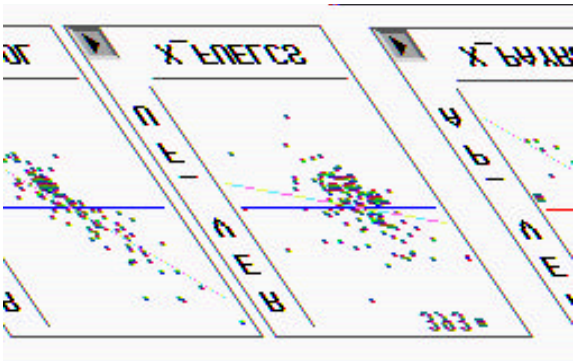
In Leverage Plots, the graphs show a horizontal line Y = 0 and a fitted regression line. This second line has an intercept of 0 and a slope equal to the parameter estimate associated with the explanatory variable in the model. The leverage plots show the changes in the residuals for the model with and without the explanatory variable. Thus, for any given data point in the plot, its residual without the explanatory variable is the vertical distance between that point and the horizontal line; and its residual with the explanatory variable is the vertical distance between the point and the fitted line. An explanatory variable having little or no effect results in a line close to the horizontal line Y = 0 -- indicating that this variable can be dropped from the model.

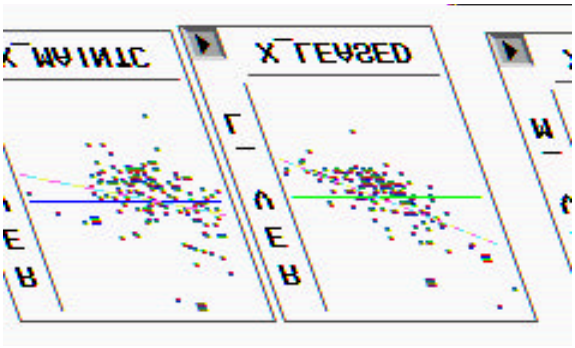
Leverage plots give increased insight if they are drawn in the scale of the original variables. SAS/INSIGHT® automatically adds in the mean of both the response and the regressor variable, and scales the plot so that the range of the original data fills the plot. If the regressor is an exact linear function of the other regressors, then the X values will completely shrink to the mean, and have no independent variation to support fitting any response variation.

Below is a set of leverage plots for the above multivariate fit. Most Analysts with whom I have worked find Leverage plots very easy to understand --***simply stated: the larger the declination of the fit line from the horizontal plane line, the more that variable influences the fit!***As can be seen, the best predictor is Payroll and the poorest is Weight.









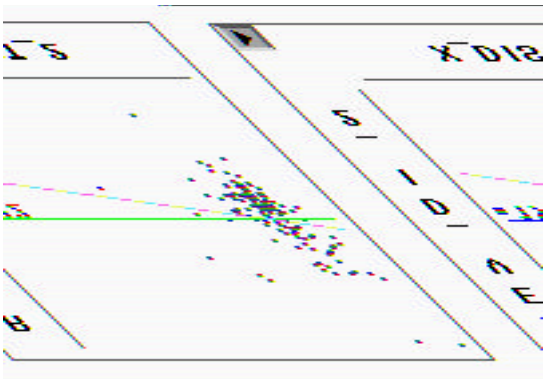
leverage is similar to our childhood experiences on a see-saw in the playground -- the further out a point is from the fulcrum, the more influence this point has on the location of the line.) Thus, the Analyst can quickly see how each of the individual points (companies) influences the declination. The highlighted points # 192 (in Distance) and # 393 (in Fuel Cost) are clearly very influential in the Leverage plots shown above. But are they wrong? As is discussed in the Introduction, we can now look at an even simpler to understand plot -- a scatterplot of Revenues vs Distance Shipped -- and use the brushing technique to help the novice Analyst understand. (Here I have highlighted point # 192 to cross-correlate our suspicions about this point.) By cross-checking the information on our survey form, we were able to confirm that a zero value had been erroneously entered for the Distance for company # 192.

Let's take two of the poorest predictors (Weight & Maintenance) out of the fit equation and look at the results:
Revenue = 1.9493 + 0.1239 Fuelcst + 0.6280 Payroll + 0.0748 Distance + 0.1560 Leasecst

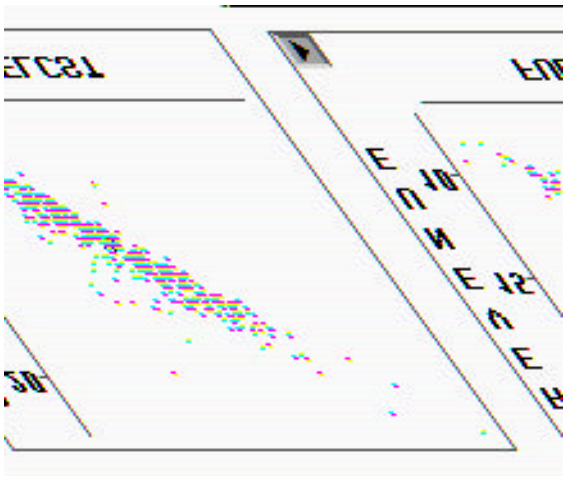
VARIABLE	Prob > F	T Statistic
Fuelcost	0.0081	0.3795
Payroll	0.0001	18.7718
Distance	0.0017	3.1887
Leasecost	0.0001	11.8256

Many Statisticians would now argue that with a maximum F of 0.0081 and all of the T Statistics above 2, that we could safely use this fit equation. However, there is an old saying, ***"In the land of the blind, a one-eyed man is king!"***

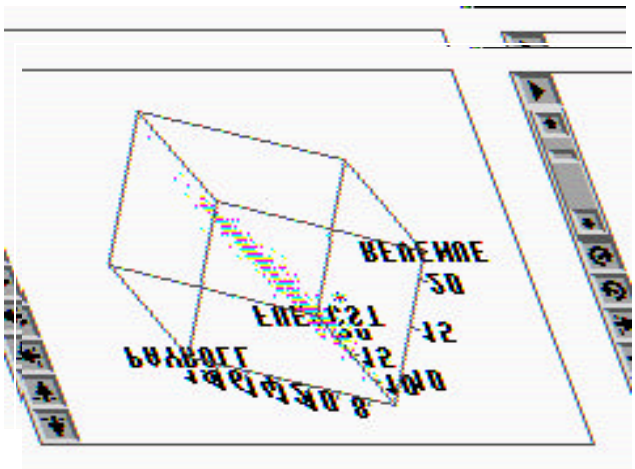
In addition to quickly being able to see declination (best predictor), the Analyst can also easily see the leverage influence of each individual company on this declination line. ***(The concept of***



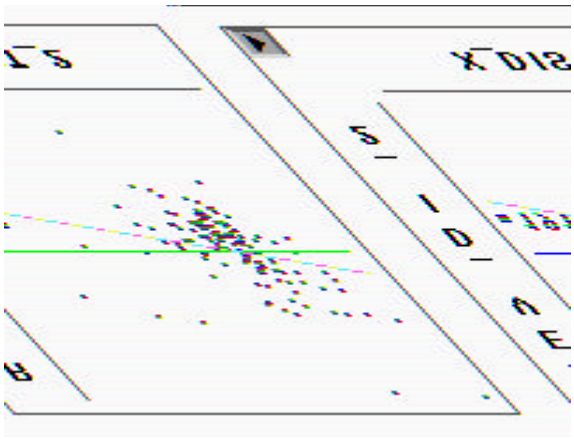
Let's take another example of how we can use Leverage plots. If we look at the Leverage plots for Payroll and Lease Costs, we note that the points are rather nicely clustered along the fit line. Let's now take the Leverage plot for Distance and replot it with just point #192 eliminated. (Before and after is shown below.) Please note how the point cloud shifts.



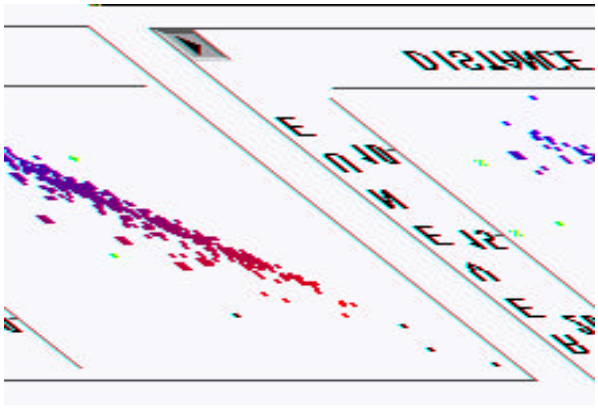
Sometimes, outlier points that are obvious in Leverage plots are not so easy to spot. In the middle of the point scatterplot cloud shown below is point # 393 (also highlighted in the Fuel Cost Leverage plot above). (Using color, Analysts can easily spot this point -- it would be shown in red against a blue point cloud.)



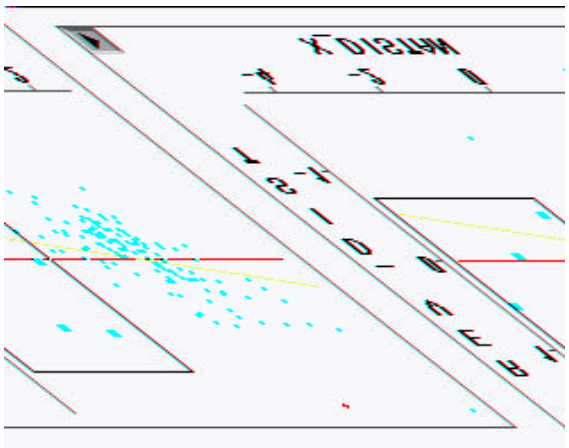
We need a 3D plot to spot this outlier. (For a more thorough discussion of outliers/inliers, readers are referred to the referenced paper by Winkler.) I have found, however, that Analysts have often had a problem understanding 3D plots at first. Often this problem comes from the poor defaults in the 3D display format. But, with a good display format and a little instruction, we have successfully added this valuable plot to our data analysis arsenal. Below, we see two plots where point # 393 is highlighted with a "*" symbol.



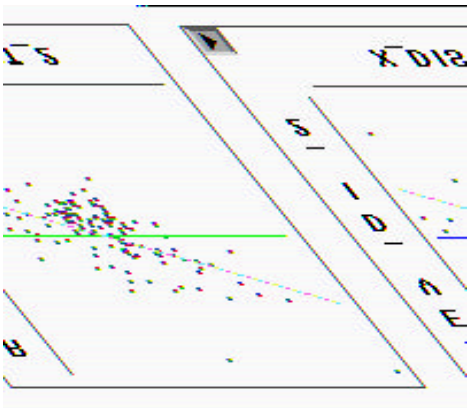
Again, we note a shift in the point cloud that give us a better distribution to these points. But, who are these companies? In the last leverage plot (above), I have highlighted some additional points that do not seem to belong. Below, these points are shown on a scatterplot of Revenue and Distance (the previously edited points # 192 and # 191 are shown on this graph by an “x”). As we can see, these points all have unusually high Revenues for the short Distances the goods were shipped -- not typical of most of the other trucking companies.



Let’s do this again with point # 191 eliminated (below).

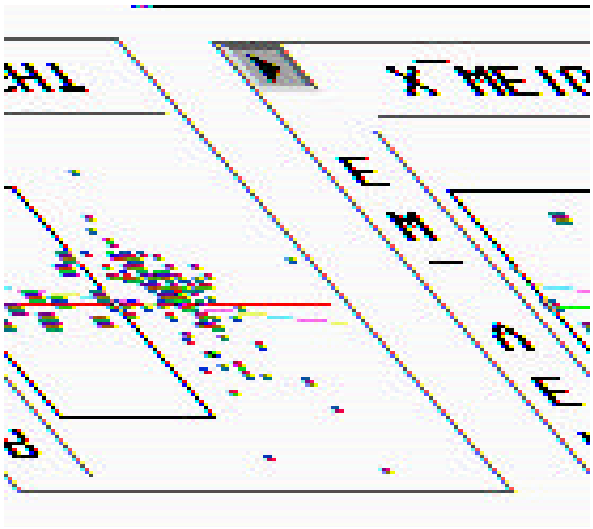


When we take this last set of companies out of the Leverage plot calculations, we get the following:



As can be seen, we now have a steeper Leverage angle, and the remaining points are also much better distributed along the fit line. (Obviously, other points also cry our for attention; again, this is a representation of a RAW, UNEDITED data set).

Finally, please also note that there is a very suspicious aggregation of strongly leveraging points (highlighted below) in the Weight Leverage plot -- these points cause the rest of the points to cluster poorly. These are the erroneous “tons/pounds” points shown earlier in this paper. Thus, paying attention to clustering in Leverage plots also helps us detect suspicious points.



So, we have seen that from even a “first pass” inspection of this data, how an Analysts can quickly spot very influential points (and point clusters) using these plots. (This even includes a multi-dimensional outlier like point # 393 noted in the first leverage plots -- an outlier that requires a 3D plot to spot -- but is readily evident with Leverage plots.) In addition, the Analyst can not only quickly conclude that the variables Shipping Weight and Maintenance Costs are the least influential variables for imputing missing/erroneous values of Revenue -- but can now not only understand WHY, and can now also quickly identify single pints and clusters of points that do not represent the majority of the other companies.

CONCLUSION:

In conclusion, the use of Leverage plots has proven to be a very useful methodology for imputing missing values. They allow our Analysts to quickly spot which are the most influential fit variables -- and help them quickly spot unusual/outlier points. In addition, we have found that these (and similar) graphical data analysis techniques have greatly enhanced our data analysis tasks -- and, most significantly, empowered our Data Analyst. To insure the understanding and implementation of these techniques, they are a key part of an on-going week-long EDA course taught by the Author to Bureau personnel.

ACKNOWLEDGMENTS:

REFERENCES

Greenberg, B.G., 1984, “SPEER, A Flexible and Interactive Editing System for Ratio Editing”, U.S. Bureau of the Census, SRD Report RR 84/18

Hogan, Howard; 1995, “How Exploratory Data Analysis is improving the way we collect business statistics”, *Proceedings of the American Statistical Association*, August 1995

Sall, J.P. (1990), “Leverage Plots for General Linear Hypotheses,” *American Statistician, 44.4*

Tukey, John W., 1970, Exploratory Data Analysis (Limited Preliminary Edition), Vol. 1, Addison-Wesley, Reading, MA.

Winkler, W.E., 1997, “Problems with Inliers”, U.S. Bureau of the Census, Washington D.C. (in press)

AUTHOR INFORMATION:

David L. DesJardins,
Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233
Tel: 301-457-4863,
E-MAIL: David.L.DesJardins@ccmail.census.gov

DISCLAIMER: This paper does not represent the position of the Bureau of the Census-- all opinions are those of the author. In addition, to insure against the possible disclosure of confidential information, only fictitious data sets were used for the illustrations in this paper.