PROBLEMS WITH INLIERS

William E. Winkler, Bureau of the Census, bwinkler@census.gov

Most edit procedures involve outlier detection and correction. They can be effective at improving the quality of certain aggregate totals obtained from a database. Inliers are fields associated with a record that are in the interior of a distribution and are in error. Inliers can arise due to systematic error and certain types of respondent and processing error. When enough records have errors due to inliers, statistical analysis of microdata can be seriously affected.

KEYWORDS: Edit, outlier, mixtures, administrative lists

1. INTRODUCTION

With the increased power and ease of use of computers, methodologists such as economists, demographers, and statisticians are no longer content to use aggregate statistical data published in tables. They prefer to use microdata because it allows them to perform policy and other statistical analyses that cannot be attempted when only aggregated data are available. By learning to run statistical packages or write elementary code to modify basic outputs or inputs, methodologists have developed both simple and sophisticated models that give much greater insight than the insights obtained from using aggregate data. Methodologists' direct analysis of databases has placed increased demands on data providers — often programmers or informaticians — to eliminate outliers and has sometimes yielded other improvements in the data or the systems providing the data.

An *outlier* is a data value which lies in the tail of the statistical distribution of a set of data values. The intuition is that outliers in the distribution of uncorrected (raw) data are more likely to be incorrect. Examples are data values that lie in the tails of the distributions of ratios of two fields (ratio edits), weighted sums of fields (linear inequality edits), and Mahalanobis distributions (multivariate normal) or outlying points to point clouds of graphs.

An *inlier* is a data value that lies in the interior of a statistical distribution and is in error. Because inliers are difficult to distinguish from good data values they are sometimes difficult to find and correct. A simple example of an inlier might be a value in a record reported in the wrong units, say degrees Fahrenheit instead of degrees Celsius.

This paper provides a number of examples of inliers and how they can seriously degrade use of microdata. In the next section, I provide background and some examples. The easiest examples are ones in which use of auxiliary data allows the usual outlier-detection methods to be applied. The third section covers how techniques for dealing with statistical mixture distributions can be used in locating and eliminating outliers. The fourth section describes how inliers and other errors can arise when groups of files are merged and the identifiers used for merging contain many errors. Such situations occur when groups of national and local health files and hospital records are merged. The final two sections are discussion and summary.

.

2. STRAIGHTFORWARD EXAMPLES

In the following subsections, we consider a variety of examples. If pairs of variables are viewed in certain ways that may be the conventional manner of looking at the variables, then we may not

be able to determine some of the pairs of variables that are in error. When we refer to a pair of variables as being in error, we mean that one or both of the variables are in error. The lack of determination is due to the fact that the erroneous records are inliers. By considering additional variables (or information), the erroneous data in certain records can be identified. Most of the identification methods are outlier-detection methods when we view the pairs of variables in different frames of reference. In the figures, inliers will usually be plotted with the symbol 'w'.

a. Inlier in distribution, outlier using prior year, or frame data

Sometimes, a respondent may make an error that affects several items in their response and does not cause the affected response to become an outlier that is more easily identified as being in error. Figure 1 illustrates the situation. Figure 1a is a plot of the true values (as if they were known to exist). Figure 1b is a plot of the observed values with some outliers included (plotted as w's). The errors in the variables were such that the inliers stayed in the interior of the distribution being plotted. An analyst can find the inliers by plotting against another data source such as the previous years values. In the resultant plot (Figure 1c), the inliers become outliers. In other words, if we change our frame of reference (by considering other variables or information), the erroneous pairs become outliers that are more easily located.

b. Geographic or industrial subcode, some subsets of companies keep records differently

If a group of records represents companies at a higher level of aggregation, it may be appropriate to consider the set of records at lower levels of aggregation. Figure 2a provides a plot of pairs of variables from a set of companies. If the records are considered in the aggregate, then it may be difficult to develop effective ratio bounds for editing. Figures 2a, 2b, and 2c show that the pairs of variables have different relationships (slopes and dispersion) on different subdomains. Erroneous pairs are denoted by 'x' in Figure 2b, by 'y' in Figure 2c, and by 'z' in Figure 2d. Figure 2e shows how the curves (which are plotted with different symbols) represent distinct subregions. To edit the data effectively (when subcategory identifiers are available), it is appropriate to develop separate sets of edits for the different subdomains. In this example, the inliers which are difficult to identify in Figures 2a and 2e are much more easily identified as outliers in Figures 2b, 2c, and 2d.

c. Large companies vs small

The example of this section is similar to the example of the previous section. Large companies may have different characteristics than small. Figure 3a is a plot of the pairs of variables from large companies, Figure 3b is the plot from small companies, and Figure 3c is the combined plot. Notice that the plot of the large companies is not as dispersed. If we use the aggregate plots (Figures 3c and 3d), then we may not locate the potential inlier in the large companies (denoted by 'w'). It may make sense to have fairly tight edit bounds on the large companies and loose bounds on the smaller ones. With the tighter edit bounds on the large companies, the inlier may become an outlier according to the usual error detection methods.

d. Pairwise detection does not work, multivariable detection does

In this example, we have three variables Z1, Z2, and Z3 that are poorly correlated with a fourth variable Y. Since they are poorly correlated (typically R-square values less than 0.25), the associated scatter plots do not provide much information about outliers. If we consider a new variables W = 2 Z1 + 2 Z2 + Z3, then we are able to find two outliers (distinguished by W's) that do not show up in Figures 4a, 4b, and 4c.

This situation arises with the Annual Survey of Manufactures. Some variables (items in a sum) are poorly correlated with a given variable and the sum of the variables is reasonably well

correlated with the given variable. The way we determine these types of situations is to produce models in which several variables are used to predict a given variable. If a group of variables is a much better predictor of the given variable than any single variable, then the group of variables can be used to develop a better edit.

3. MIXTURES

In some situations, a set of pairs of variables from records may be the mixture of two different distributions. High overlap of the point clouds may make it difficult to decide what the two mixtures are and how to find points and correct them. If auxiliary information is available (as illustrated in section 2), then the auxiliary information may allow finding wrong records. Figure 5 illustrates a situation in which most companies may have accurate information associated with them and a few smaller companies have erroneous information (inliers) that may arise somewhat systematically. Such errors may correspond to companies reporting quantities in which certain taxes are not added in properly and in which they have subtracted an inventory item that should not be subtracted. Figure 5a is a scatterplot of pairs of variables associated with companies that have accurate quantities associated with them. The slope of the regression line is 1.0 and the R-square value is 0.72. Figure 5b is the corresponding scatterplot of smaller companies having erroneous pairs of variables. Figure 5c is the combined plot. The slope is 1.3 and the associated R-square value is 0.65.

Whether a given scatterplot corresponds to a mixture may not be easy to determine. In some situations, viewing the point cloud may give sufficient information. In other situations, it may be necessary to fit specific mixture models (e.g., Titterington et al 1985). This example shows how analyses can be in error if many inliers are in the microdata. The mixture model analyses may allow a correct regression analysis using the variables but it may not yield corrected microdata.

4. ADMINISTRATIVE LISTS - ANALYTIC LINKING

Appropriate microdata can support valid statistical analyses. Some microdata reside in multiple files for which common identifiers are unavailable. Use of aggregate data does not allow evaluation of correlations of data across files and accurate analyses on subdomains. We are interested in an edit/imputation methodology for use on two or more files. In many situations common information is not uniquely identifying and is subject to significant error. This makes linking (computer matching) of the correct corresponding microdata records very difficult. In a subset of these situations, each source file may contain noncommon quantitative data that can be connected with appropriate edit/imputation models. For instance, we might want to use files of businesses that only have difficult-to-use name and address information in common. One file might contain the energy products consumed by the companies. The other file might contain the types and amounts of goods they produce. Another situation could arise with files on individuals in which one file has income data, another information about health-related products, and a third information about receipts of supplemental payments. If we have an appropriate methodology, then we may be able to edit/impute the microdata. As with selective editing, a primary goal may be to produce valid aggregate estimates or a few elementary statistical analyses. If resources are available and even more powerful techniques are available, we may be able to produce valid microdata files.

For simplicity, we consider the situation where records in a file are formed by merging (matching) two files. For each record, we assume that a matching score is available that

summarizes the quality of the matching process. We assume that there is an unobserved indicator function that tells whether the record is in error or not. Note that a similar scenario could also arise when records contain certain types of systematic error. In the following example, name and address information is used in creating a file representing the merger of two other files. One file contains the y-variable and the other file contains the x-variable.

Figure 6a illustrates the ideal situation in which we have perfect matching information for bringing files together. Figure 6b shows the observed data as a result of a matching process in which the identifiers used in matching contain severe errors. It is difficult to determine which points correspond to pairs of variables that should be brought together. Pairs of variables that are erroneous (inliers) because of incorrect linking are denoted by 'o'. The inliers almost overwhelm the correct data. With a small subset of pairs representing having high matching scores (about 5% of all pairs), we can develop a model for the data relationships and determine some pairs that may be outliers. The modeling is not entirely accurate because more than 10% of the 100 pairs in the small subset are associated with erroneous linking and the subset is quite small for prediction purposes. Figure 6c illustrates the situation in which we replace some pairs with other pairs in which the y-variable is predicted using a crude regression model. We add the predicted y-variables to the file that only contains the x-variables. This yields an additional matching variable which we use to improve matching during a second pass.

Figure 7a illustrates the true situation among the pairs if we had perfect matching information. Figure 7b shows that the variables brought together during the second matching pass are more closely related. The reason that the plot in the right-hand column contains fewer points is that the predicted y-variables improve linking sufficiently to substantially reduce erroneous links (inliers). Finally, Figure 7c shows the result of some final modeling in which outliers are removed.

5. DISCUSSION

To deal with inliers in most situations, we convert them to outliers using additional information that may be available in the files being edited. Only in the last two examples, mixtures and administrative lists-analytic linking, do we need to perform more work to create additional information that allows us to improve the edit/imputation process.

6. SUMMARY

This paper describes methods of dealing with inliers (erroneous variables that are in the interior of a distribution) and not always easily located.

REFERENCES

Scheuren, F. and Winkler, W. E. (1996), "Recursive Merging and Analysis of Administrative Lists and Data," *American Statistical Association, Proceedings of the Section on Government Statistics,* 123-128 (available on http://www.amstat.org in section on govt statistics).

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, John Wiley: New York.

Figure 1a. Current Year – Truth
Two values from current year

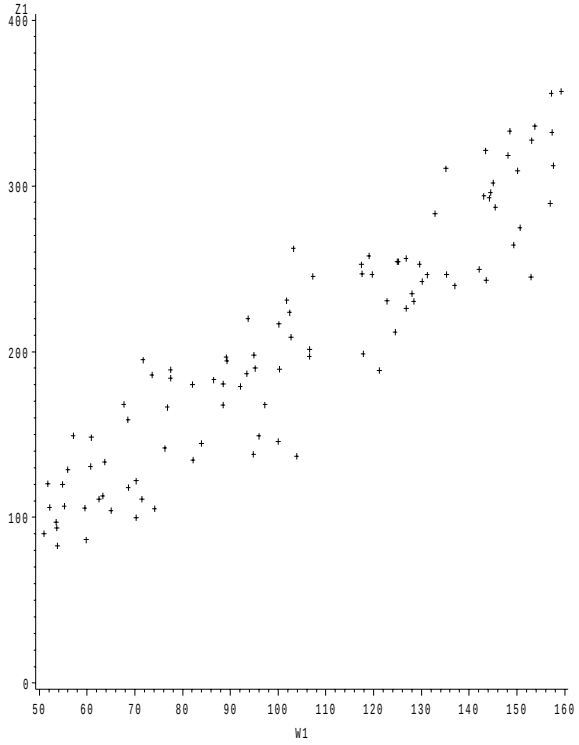

Figure 1b. Current Year – Observed Errors
Two values from current year

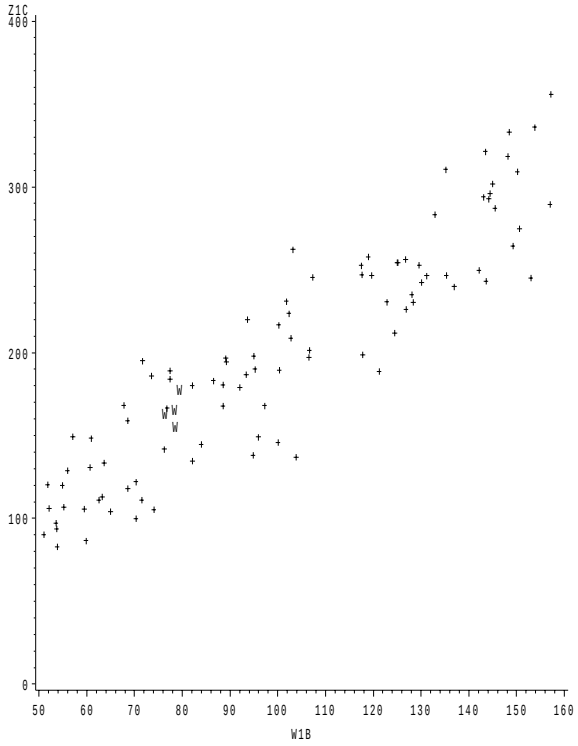

Figure 1c. Previous Year against Current Year
Corresponding quantities

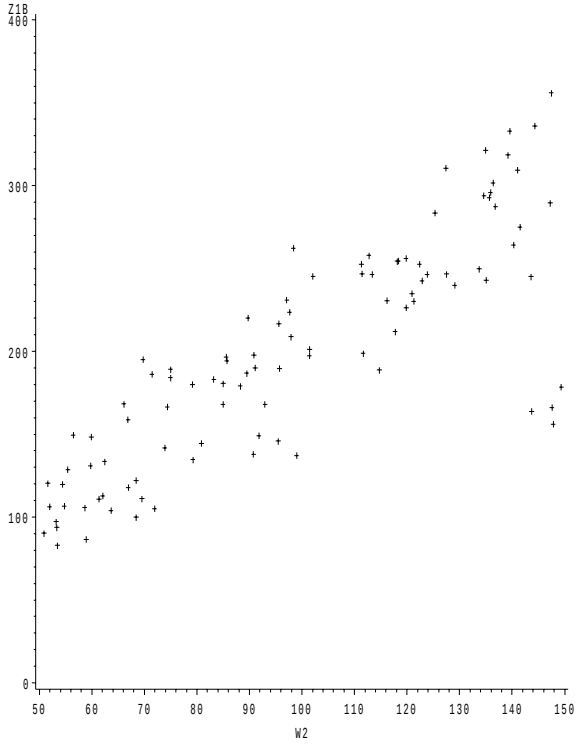

Figure 1d. Different Symbols – Inliers Shown

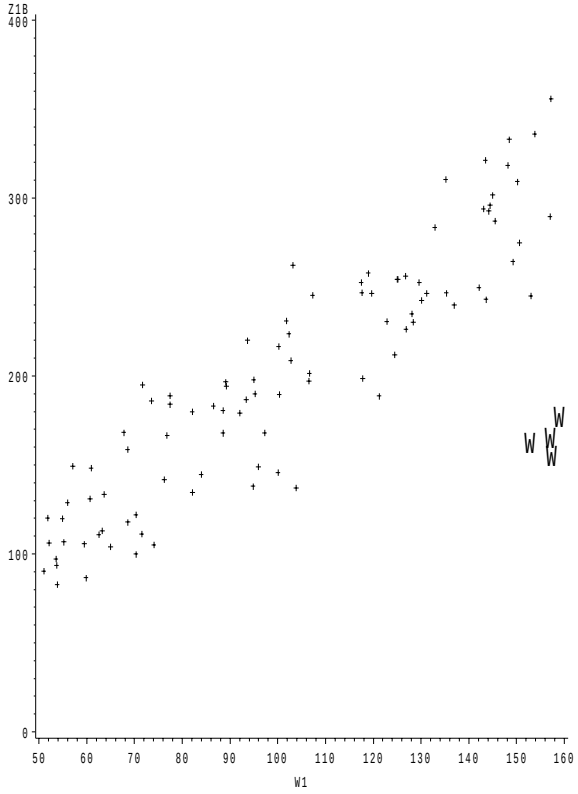

Figure 2a. Mixture of Three Curves

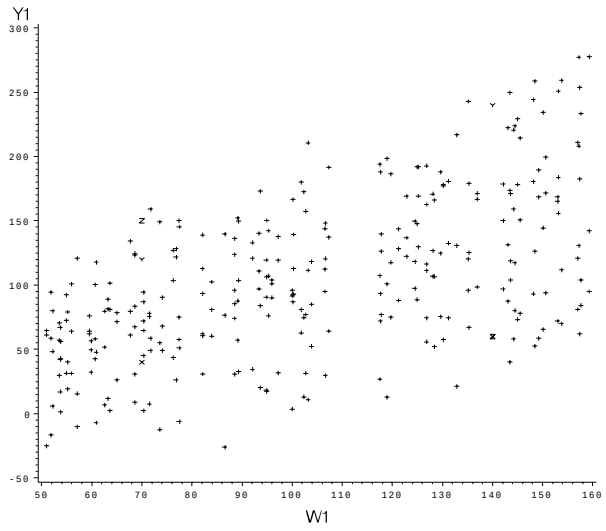

Figure 2b. First Original Curve

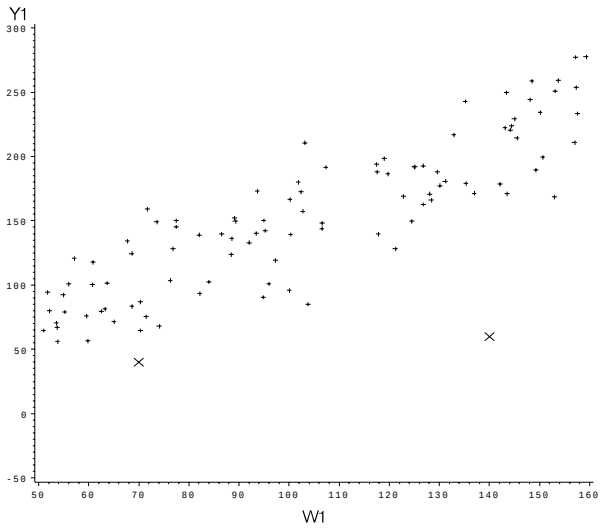

Figure 2c. Second Original Curve

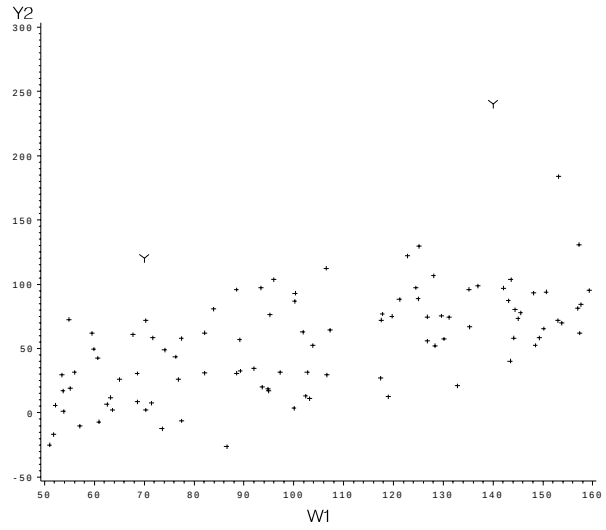

Figure 2d. Third Original Curve

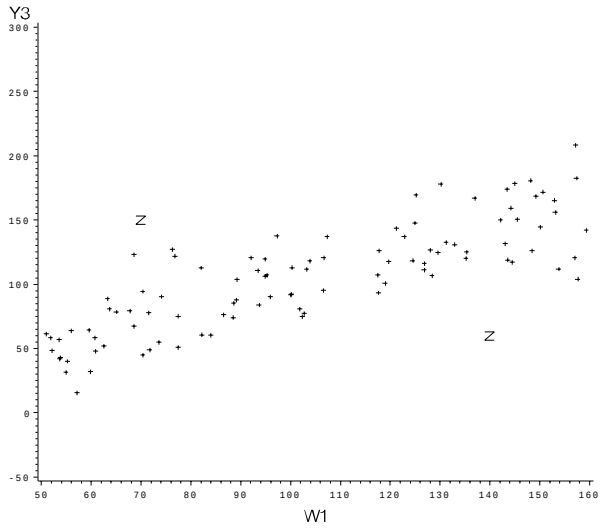

Figure 2e. Three Curves — Inliers Shown
Separate Ratio Edits Needed

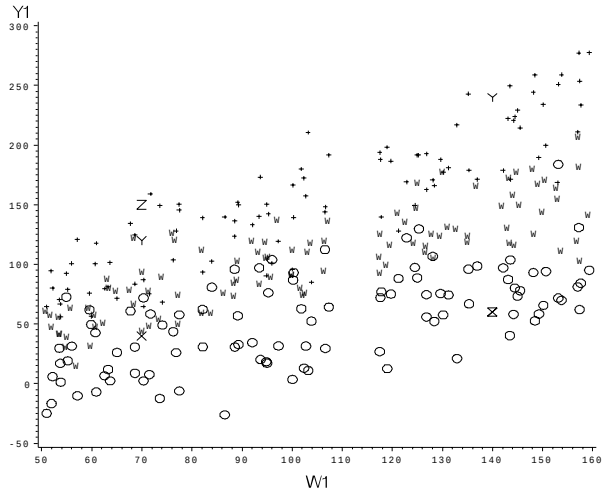

Figure 3a. Original Curve – Large

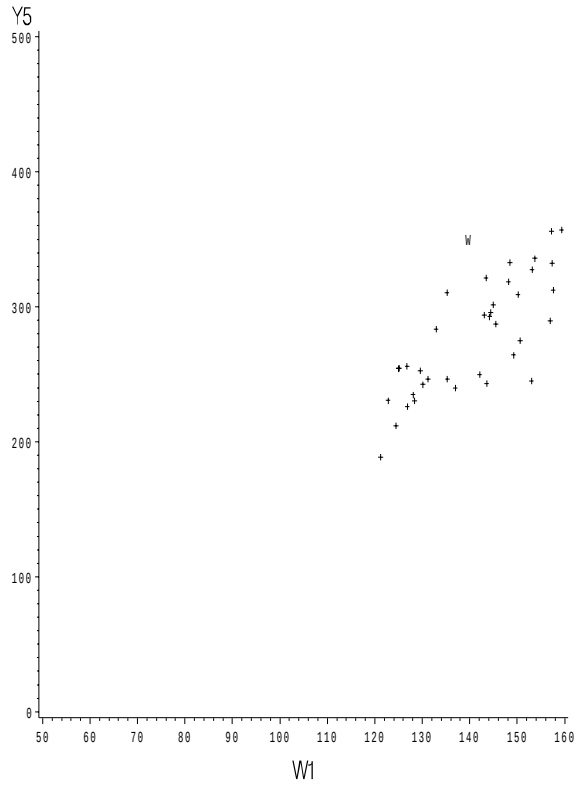

Figure 3b. Original Curve – Small

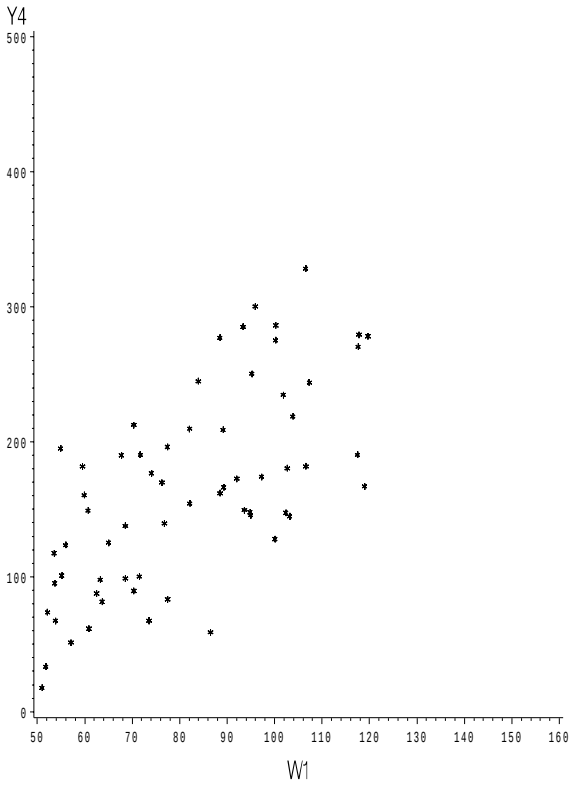

Figure 3c. Mixture of Large and Small

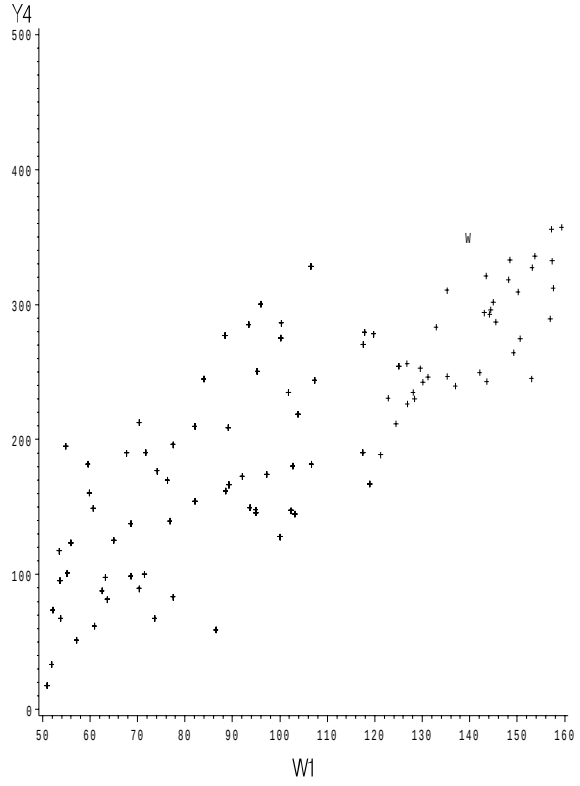

Figure 3d. Large and Small– Distinguished Separate Ratio Edits Needed

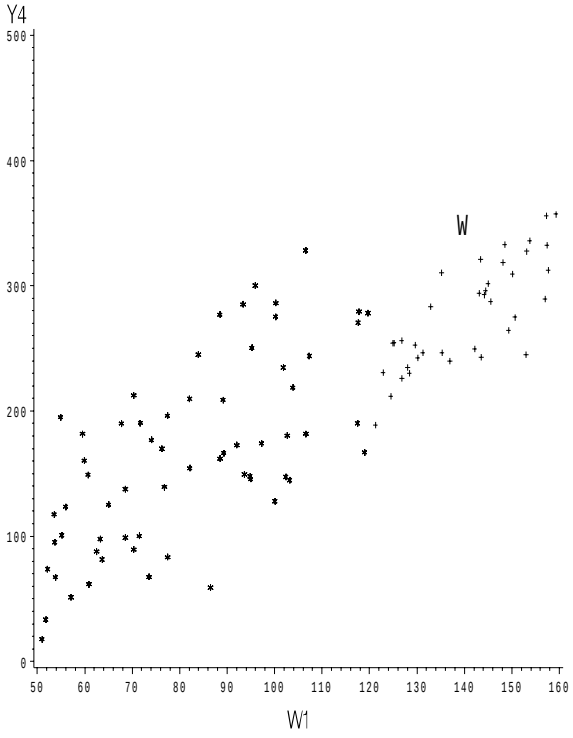

Figure 4a. First of Three Curves
Poor Prediction for Ratios, Y vs Z1

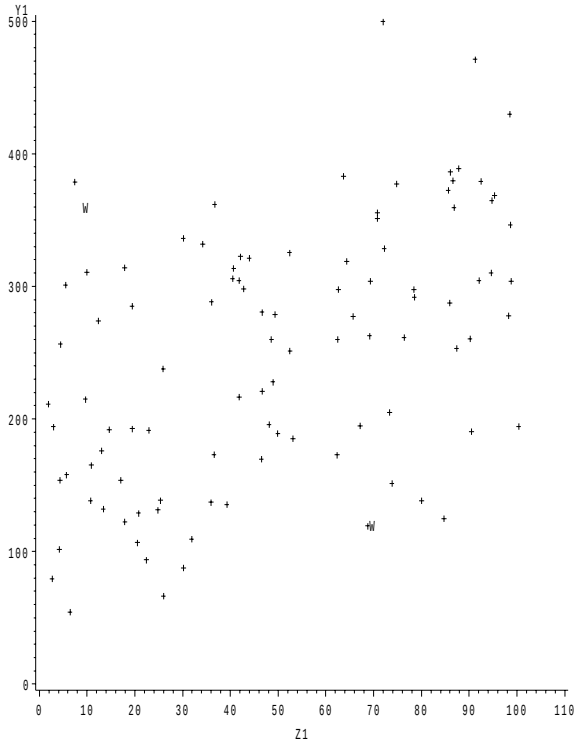

Figure 4b. Second of Three Curves
Poor Prediction for Ratios, Y vs Z2

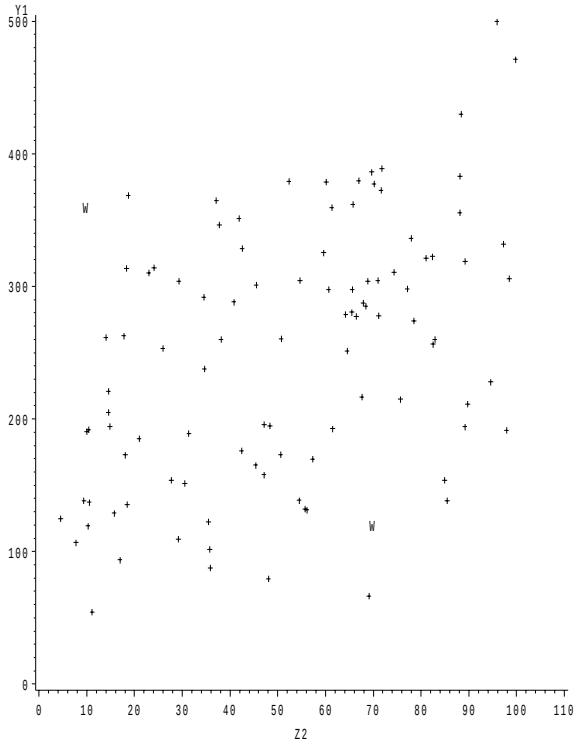

Figure 4c. Third of Three Curves
Poor Prediction for Ratios, Y vs Z3

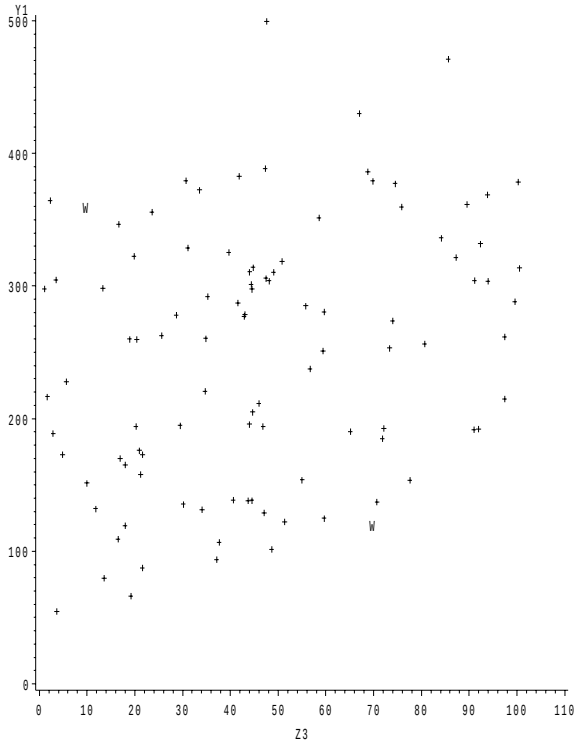

Figure 4d. Good Predictor Curve
W = 2 Z1 + 2 Z2 + Z3
Inliers Distinguished, Y vs W

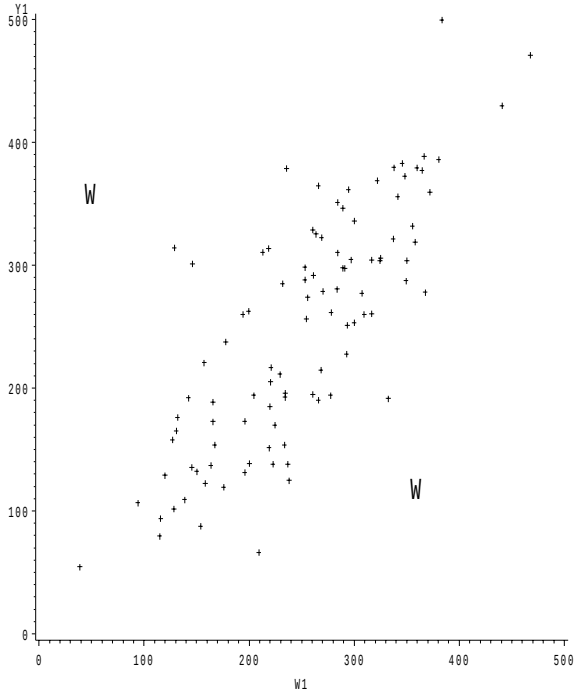

Figure 5a. Most Companies
slope = 1.0, R – square = 0.72

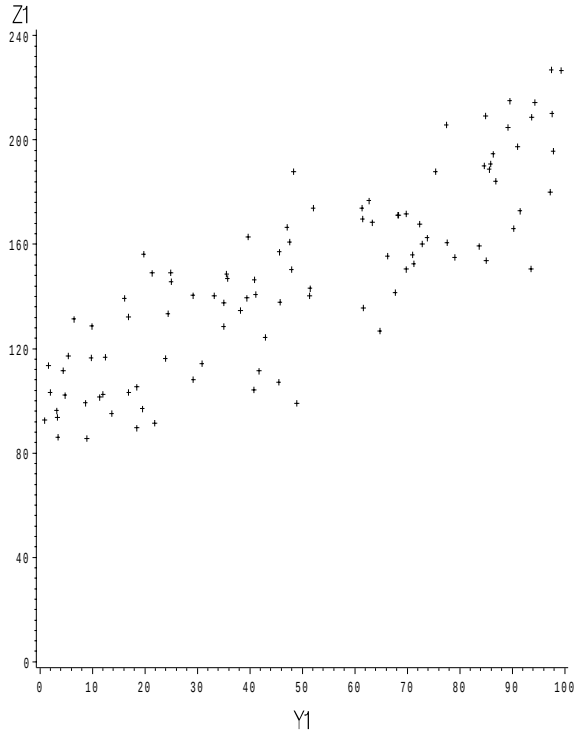

Figure 5b. Smaller Companies
Systematic Error

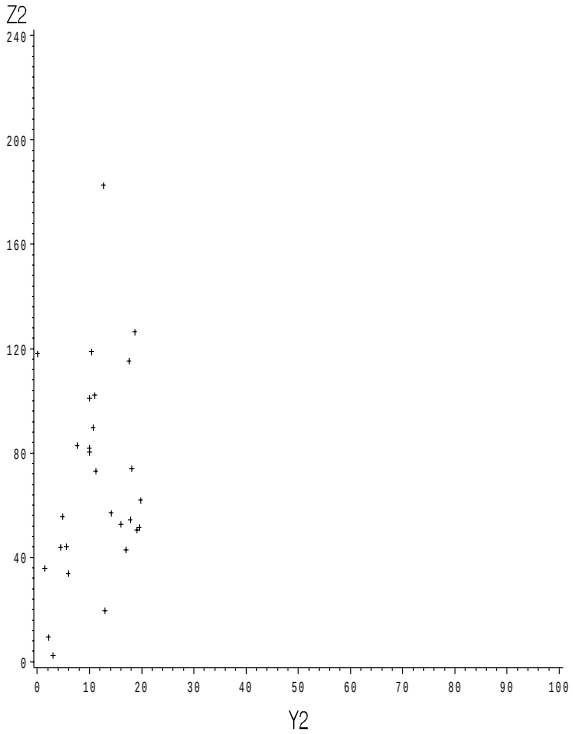

Figure 5c. Mixture of All Companies
Use Mixture Methods to Distinguish
slope = 1.3, R – square = 0.65

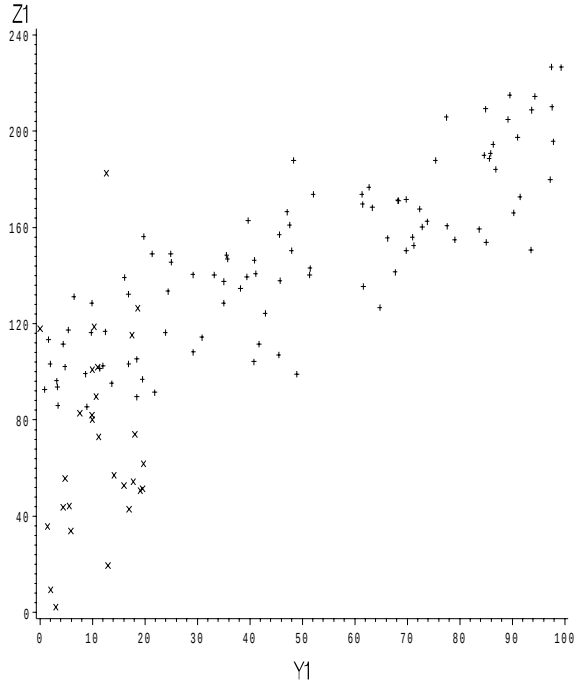

Figure 6a. 1st Pass Matching, True Data

1104 Points, beta = 5.85, R-square = 0.43

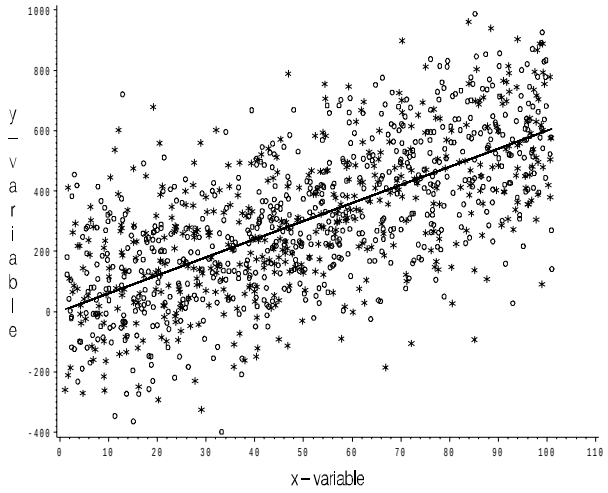

Figure 6b. 1st Pass Matching, Observed Data

1104 Points, beta = 2.47, R-square = 0.07

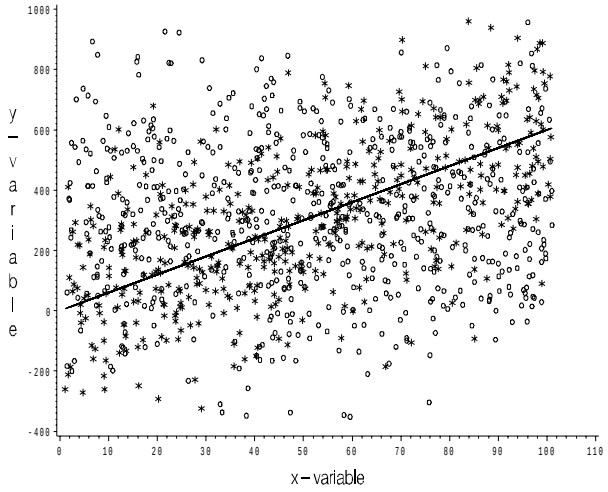

Figure 6c. 1st Pass Matching, Outlier-Adjusted Data

1104 Points, beta = 4.78, R-square = 0.40

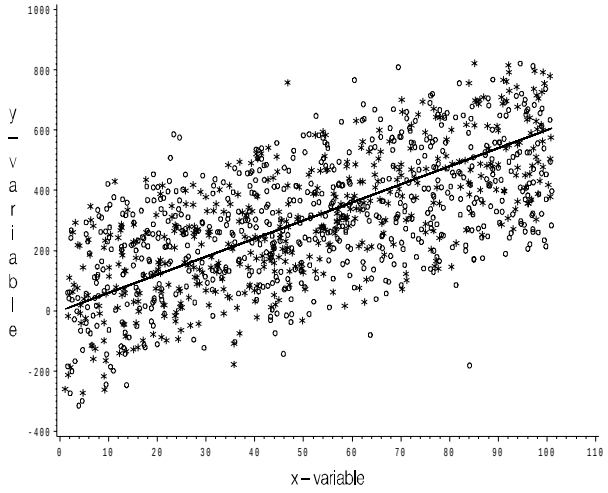

Figure 7a. 2nd Pass Matching, True Data

650 Points, beta = 5.91, R-square = 0.48

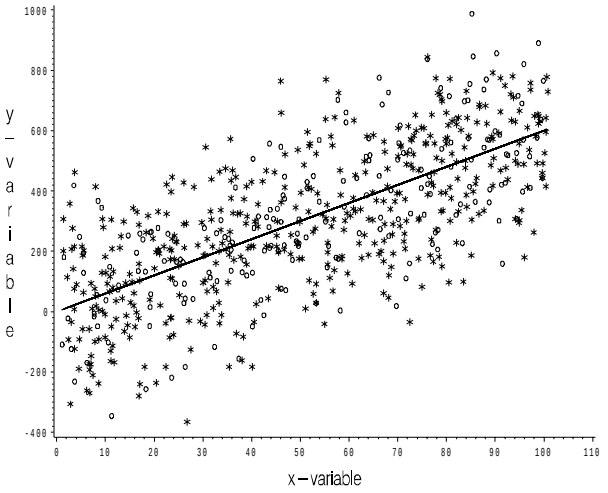

Figure 7b. 2nd Pass Matching, Observed Data

650 Points, beta = 4.75, R-square = 0.33

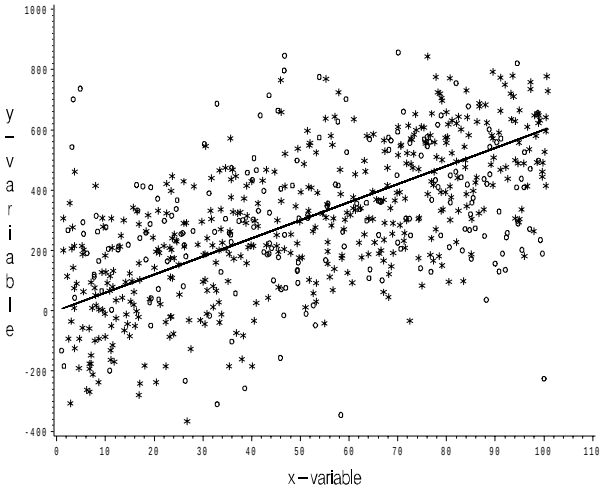

Figure 7c. 1st Pass Matching, Outlier-Adjusted Data

650 Points, beta = 5.26, R-square = 0.47