Study of Extreme Values and Influential Observations[1]

Mark E. Asiala & Alfredo Navarro, US Census Bureau
Mark E. Asiala, 2403-2 DSSD, Washington DC 20233-7613

Key Words: extreme values, outlier detection, design effect, generalized variance

1 Introduction

In our development of the plans for the 1998 Dress Rehearsal Long-Form Generalized Variance, we looked to employ a new outlier detection methodology that could automate the adhoc process used in 1990. The generalized variance program generates a group design effect for 60 different data item groups by calculating a weighted average of the design effect for each data item contained in that group over each Final Weighting Area (FWA) in the Dress Rehearsal. In 1990, the process of identifying outliers was done using a combination of graphs and studying the relative absolute deviation. It was desired that for the 1998 Dress Rehearsal plans a more automated and objective means could be used.

In this summary, we present two major approaches to this problem. The first major approach examines the data to determine extreme values using different estimators of scale and then down weights acccordingly. Some general background for comparing different estimators is given before a comparison of four leading candidates for detecting extreme values is presented. The second major approach examines the influence of each of the observations and applies a down weighting based on the influence of the observation rather than its raw value. Each of these methods is tested on a dataset obtained from the 1998 American Community Survey. We conclude with a few remarks about our study.

2 Background and Description of 1990 Methodology

The long form contains a number of housing and population questions which are summarized for public release. Each estimate released to the public must have a corresponding estimate for the standard error (SE). Since the summary tables released contain a large number of estimates, SEs are published using the generalized variance method of multiplying the SE obtained if a simple random sample was used by an appropriate design effect which accounts for the cluster sampling effects (by household) and the intra-cluster association. Users are given tables and the formulas used to produce the tables in order to make the necessary calculations. Design factors are published for 60 different groups of housing and population characteristics and users must select the appropriate group design effect for their particular estimate. For example, a user estimating the SE for the number of children under 17 in poverty would use the group design effect for poverty in their calculation.

In 1990, the method for detecting extreme values centered around calculating the relative absolute deviation of the Actual SE from the Predicted SE using the group design effect for each observation. The relative absolute deviation (RAD) was defined as

$$RAD = \frac{|\text{Actual } SE - \text{Predicted } SE|}{\text{Predicted } SE} \times 100\%$$

Observations for which the RAD was greater than 40 percent were flagged. Next they looked at a graph of Predicted SE vs Actual SE and used the graph to determine which of the flagged observations would be designated as an extreme observation. Extreme observations were removed from the group design effect calculation.

3 Detecting Extreme Values

Four methods for detecting extreme values were applied and evaluated in our work. Each of these was based on a different estimator of scale taken and/or adapted from Rousseeuw & Croux (1993). To place these estimators in context, we first review some properties of estimators of scale.

3.1 Properties of Estimators of Scale

The topic of theoretical properties of estimators can be found in many standard textbooks on robust statistics including Huber (1981), among others. We

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

concentrate here on four main properties which we found useful for studying estimators of scale.

3.1.1 Breakdown Point

The first property to be defined is the breakdown point. We define the (explosion) breakdown point as the minimum percentage of the data which, when replaced by some arbitrary values, makes the estimate equal to infinity. Some examples for estimators of scale include the standard deviation and the average absolute deviation, which have a breakdown point of zero-percent. This contrasts to the interquartile range which has a breakdown point of 25%. For location estimators, the mean has the same breakdown point as the standard deviation and the median has a breakdown point of 50%. We would expect that estimators with higher breakdown points would be less affected by high-valued outliers since it takes a greater portion of the data to significantly affect the value of the estimator. For example, the standard deviation can be greatly increased by the presence of one extreme value whereas the inter-quartile range (with a higher breakdown point) is insensitive to the presence of outliers until they affect the first or third quartiles (which is why its breakdown point is 25%). The maximum breakdown point possible for an estimator is 50% so this gives a mark to measure all esimators of scale against.

3.1.2 Influence Function

A second measure is given by the influence function (or influence curve) and the gross-error sensitivity. The influence function for a distribution function F and a functional T is defined as

$$IF(x,T,F) = \lim_{s\to 0} \frac{T((1-s)F + s\delta_x) - T(F)}{s}$$

where δ_x denotes the pointmass observation 1 at x. This function can be interpreted as a type of derivative which measures the change in the estimate due to a small increase of data at a position x. While one can directly compare influence functions of different estimators, it is often sufficient to look at the maximum of the influence function over the dataset (or the infimum if applying to a theoretical infinite set). This provides a one-number comparison called the gross-error sensitivity defined as

$$\gamma^*(T,F) = \max_x |IF(x,T,F)|$$

Thus we see that an estimator with a higher gross-error sensitivity will be more greatly affected by the presence of outliers than one with a lower gross-error sensitivity.

3.1.3 Efficiency

A third property is the variance or efficiency of the estimator. One can measure the efficiency of an estimator of scale, for example, by comparing the variance of the estimator to that of the standard deviation over a Gaussian distribution. Low efficiency implies high variance for the estimator and thus it is desirable to have an estimator with high efficiency. By definition, the standard deviation is 100% efficient.

3.1.4 Relation to Symmetry

The last property is the appropriateness of some estimators for symmetric versus asymmetric data. The classic examples are for location estimators, the mean versus the median, and for scale estimators, the standard deviation versus the use of quartiles. We will see that in the application of estimators of scale to outlier detection, the use of a particular location estimator or the use of no location estimator can be very important since outliers may not be situated at equal distances from the location estimator if the data are asymmetric.

3.2 Four Estimators of Scale

At the base of each outlier detection method is a different estimator of scale. As stated earlier, methods such as standard-deviation or the average absolute deviation are subject to instability because of their low breakdown point. All four of these estimators are based instead on either the median or an order statistic.

3.2.1 Median Absolute Deviation

The first estimator of scale is the median absolute deviation (MAD_n). It is equivalent to the average absolute deviation except using the median rather than the mean. In symbols it is defined

$$MAD_n = 1.4826\, b_n \operatorname*{wmed}_{i=1,\ldots,n} \left| x_i - \operatorname*{wmed}_{j=1,\ldots,n} x_j \right|$$

where b_n is a finite population correction parameter, wmed is the weighted median, and the coefficient 1.4826 is for consistency with the standard deviation over normal distributions. The weighted median is calculated by first sorting the observations and then finding the observation where the sum of the weights below the observation is equal to the sum of the weights above the observation. The use of the median in place of the average improves the estimator in several ways. The breakdown point of the MAD_n is 50%, the maximum possible, improved from 0% for the average absolute deviation or the standard deviation. It also has a bounded influence function which yields a finite value for the gross-error sensitivity. In fact, it can be shown that the MAD_n has the best gross-error sensitivity for symmetric data but it does not fare as well for non-symmetric data. The Gaussian efficiency, however, is fairly low at 37% compared to the median's location estimator efficiency of 64%.

Technically, the MAD_n is also easy to calculate with computation times on the same order as the number of observations and the storage required is also on the same order as the number of observations. Overall, this means that for symmetric data it is very hard to beat the MAD_n on all counts except for its efficiency.

3.2.2 S_n (unweighted and weighted)

S_n was designed to improve over MAD_n for asymmetric data. The problem MAD_n has for asymmetric data is that it counts observations equally for values which are a fixed distance above or below the median. For asymmetric data, however, outliers may not lie symmetrically about the median. S_n thus has no location estimator in its definition in order to avoid biasing itself against asymmetric data. Its definition is given by

$$S_n = 1.1926\, c_n \operatorname*{lomed}_{i=1,\ldots,n} \operatorname*{lomed}_{j \neq i} |x_i - x_j|$$

where c_n is a finite population correction parameter and the coefficient 1.1926 is for consistency with the standard deviation over a Gaussian distribution. Note that the inner low median excludes j equal to i.

S_n has a 50% breakdown point like MAD_n and also has a bounded influence function. We already stated that MAD_n has the best gross-error sensitivity for symmetric data but S_n performs better than MAD_n for asymmetric data. The improvements for asymmetric data do not prevent S_n from performing well for symmetric data with S_n only slightly behind MAD_n. One final improvement over MAD_n is that the Gaussian efficiency for S_n is 58% compared to the 37% efficiency of MAD_n. So S_n has made significant improvements on two major grounds. The improvements do not come for free as S_n requires more computation time than MAD_n. With careful programming however, the calculation of S_n takes only on the order of $n \log n$ time (n=number of observations) as compared to order n time for MAD_n. The amount of storage is comparable to MAD_n at order n size.

A weighted version of S_n was proposed using the following definition,

$$S_n(weighted) = 1.1926\, c_n \operatorname*{wlomed}_{i=1,\ldots,n} \operatorname*{wlomed}_{j \neq i} |x_i - x_j|$$

where wlomed is the weighted low median. It is the weighted equivalent of the low median and is calculated in a manner similar to the weighted median.

3.2.3 Q_n

Q_n is one more attempt to improve upon both MAD_n and S_n. There is still room for improvement in the Gaussian efficiency compared to S_n. So Q_n replaces the double median of S_n with an approximate first quartile calculation. In symbols it is defined as

$$Q_n = 2.2219\, d_n \{|x_i - x_j| : i < j\}_{(k)}, \quad k = \binom{h}{2},$$

where (k) is the k^{th} order statistic, h is defined as $h = \lfloor \frac{n}{2} \rfloor + 1$, d_n is a finite population parameter and the coefficient 2.2219 is for consistency with the standard deviation over a Gaussian distribution.

Q_n keeps the advantages of S_n with a 50% breakdown point, a bounded influence function, and improved performance for asymmetric distributions. The key improvement for Q_n is the increasing of the Gaussian asymptotic efficiency to 82% which is the best of the four estimators. This benefit is not realized, however, until n is larger than approximately 50. Its gross-error sensitivity is larger than S_n by a small amount. Q_n also takes an equivalent amount of processing time and storage space as S_n.

Since the main advantage of Q_n, efficiency, does not appear until n is greater than 50 it is recommended that S_n is used for smaller datasets. In addition, S_n's lower gross-error sensitivity may make it better for detecting outliers since Q_n offers no offsetting advantages for this size dataset. For datasets larger than 50, Q_n's greater efficiency can offset its higher gross-error sensitivity and make it the prime choice.

3.3 Application to Outlier Detection

The application of each of these estimators of scale to outlier detection involves a standardization of observation values. These standardized values are then compared to a fixed value. Since the estimators of scale were normed to the standard deviation for a normal distribution, a value of 2.5 or 3.0 may be used in order to determine outliers. It is our intention, however, to empircally determine the best comparison value by evaluating which value produces the best results based on a number of criteria. It is expected that this value will fall between 2.5 and 3.5. The four test statistics using MAD_n, Q_n, and S_n (unweighted and weighted) respectively are

1. $\frac{|x_i - \mathrm{wmed}_j\, x_j|}{MAD_n}$
2. $\frac{\mathrm{lomed}_{j \neq i}\, |x_i - x_j|}{Q_n}$
3. $\frac{\mathrm{lomed}_{j \neq i}\, |x_i - x_j|}{S_n}$
4. $\frac{\mathrm{wlomed}_{j \neq i}\, |x_i - x_j|}{S_n(weighted)}$

We note that given the similarity of S_n to Q_n, we are able to use the same numerator for both test statistics.

4 Studying Influence Rather Than Raw Data

Another method which was explored was making the influence a single observation has on the group design effect the object of study rather than the raw data values. This has the advantage that it includes the weight of an observation as well as its departure from the group design effect in a single number. Extreme observations with little weight are thus treated less harshly than extreme observations of higher weight. Belin, Schenker, and Zaslavsky (1999) outline a method they employed using the 1990 PES data and compared their method to other standard practices being used to deal with influential observations.

Their method as applied to our situation involves five basic steps:

1. Compute the influence for each observation on the group design factor.
2. Calculate the location/scale estimates for the values of the influence function.
3. Calculate the down weights using the location and scale estimates found in (2).
   (a) Match a t-distribution to the values of the influence function using a QQ-plot in order to obtain a degrees of freedom estimate ν.
   (b) Standardize the observations using the location/scale estimates found in (2) creating a new set of observations z_i.
   (c) Calculate the down-weighting factors using the following formula
   $$wgt_i = (1 + z_i^2/\nu)^{(-1)}$$
4. Multiply the original weights of the raw data by the weights found in (3c).
5. Use the weights calculated in (4) to calculate a new group design effect.

This method was applied to our situation of calculating the group design factors. To calculate the influence for an observation we begin with the equation for the group design effect, DF_g, modified to contain an inclusion parameter, I_t, which is equal to one if the observation is included and zero otherwise.

$$DF_g = \frac{\sum_{t=1}^{n} Est_t\, DF_t\, I_t}{\sum_{t=1}^{n} Est_t\, I_t}$$

where Est_t is the estimated count of the t^{th} data item and DF_t is the design effect of the t^{th} data item. We note that if all the I_t's are equal to 1 then we have just the standard equation for the weighted average. We next differentiate the above with respect to I_k to approximate the influence that the inclusion of the k^{th} data item has on the group design effect.

$$Infl(k) = \frac{\partial DF_g}{\partial I_k} = \frac{Est_k}{\sum_{t=1}^{n} Est_t I_t}(DF_k - DF_g)$$

Using the fact that the I_t's are equal to one,

$$Infl(k) = \frac{\partial DF_g}{\partial I_k} = \frac{Est_k}{\sum_{t=1}^{n} Est_t}(DF_k - DF_g)$$

We note that the RAD definition given in Section 2 can be simplified using two definitions: Actual $SE = DF_t(SRS\ SE)$, Predicted $SE = DF_g(SRS\ SE)$ where $SRS\ SE$ is the SE obtained if one assumed a simple random sample. This transforms the RAD definition as applied to our case to

$$RAD(DF_t) = \frac{|DF_t - DF_g|}{DF_g} \times 100\%$$

Thus the influence of a data point is directly proportional to the weighted value of the RAD (using the estimates as weights) which was studied in 1990. This made this approach look promising. The results of both major approaches (extreme value detection on raw values and studying of influence values) on an actual dataset are discussed in the next section.

5 Application of the Methodologies to a Sample Dataset

The data set comes from the 1998 American Community Survey Generalized Variances which utilizes the same methodology as the 1990 Census Long Form Generalized Variances. The data set includes item, state, county, tract, data group, data item, replicate SE, $SRS\ SE$, and the data item/tract-level design effect. The total data came from 9 sites.

5.1 Application of RAD, $\mathbf{MAD_n}$, $\mathbf{S_n}$, and $\mathbf{Q_n}$

Each of the four outlier detection schemes were applied to the design factors. Since the group design effect is a weighted average, an attempt was made to create programs to calculate a weighted version of each of the three estimators of scale. This was done for both MAD_n and S_n but was not done for Q_n due to some programming difficulties. In our final comparison we look at a weighted MAD_n and S_n and also at a unweighted (normal) Q_n and S_n.

From the original set of data, we selected only those tracts which had non-zero estimates for that data item and whose design effects were defined. This dataset contained a total of 803,444 data item / tract combinations for the nine sites. Each of the four methods for determining extreme (raw) values were employed using cutoff values ranging from 2.5 to 3.5 by 0.1 increments. The recalculated group design effects were compared by average RAD value, weighted average RAD value using the old weights and the recalculated weights, and also by the median RAD value. The cutoff for each method which produced the best results was then kept.

5.1.1 Differences in number of outliers

- For our data we used the following cutoffs for MAD_n, Q_n, S_n, and weighted S_n: 3.2, 3.0, 3.0, and 3.1. This resulted in 7406, 55967, 51799, and 15053 observations being identified to be down weighted respectively.

- Using a cutoff value of 40%, *RAD* would identify 205,043 extreme values which would need graphical followup. Using a weaker cutoff value of 60%, *RAD* would identify 57,707 extreme observations requiring followup.

- In most cases the data was not symmetric for each group. Typically, the data was skewed right with a longer tail towards higher values. This resulted in Q_n and both versions of S_n identifying extreme values mainly in the tail whereas MAD_n and *RAD* identified extreme values symmetrically about the group design effect.

5.1.2 Computation Time

There were some significant differences in the computation time for the various outlier detection methods. Using all sites in one file, all four methods were run simultaneously on a Compaq Alpha. The times were 20 min, 30.8 min, 30.4 min, and 57.9 min for the weighted MAD_n, unweighted Q_n, unweighted S_n, and weighted S_n programs respectively.

The times are consistent with the speed of the algorithms published in the original paper by Croux and Rousseeuw (1992). The most noteworthy result is the increase in time from the unweighted to the weighted version of S_n. This is mainly due to the more costly algorithm for performing the weighted high median used in calculating the weighted S_n versus the more efficient algorithm for S_n that Croux and Rousseeuw give in their paper. This makes the weighted S_n rather costly for computation requirements.

5.2 Application of Influential Observation Method

Using the formula derived in the previous section, the influence of each observation on the group design effect was calculated. We calculated the mean and standard deviation of the values of the influence function and plotted those on a QQ-plot of standardized values versus quantiles of the t-distribution for $n = 0, 1, 2, 4, \ldots, 32$ degrees of freedom as well as standardized values versus quantiles of the normal distribution. From these plots we determined that the best fit came from a t-distribution with $n = 2$ degrees of freedom. This was then plugged into the equation for the weight modifier and used to down weight all the observations. A new group design effect was calculated using the new weights.

6 Recalculation of Group Design Effects

We had five different methods of identifying/dealing with either extreme values or influential observations: MAD_n, Q_n, S_n unweighted, S_n weighted, and the influential observation method. The last method, which is based on a smooth down weighting of influential observations, has a method of down weighting that is built-in. The other methodologies simply identify the extreme values. What is done with those observations is a separate decision. Clearly, one faces three choices: eliminate the observations (set weights equal to zero), ignore the information and include them fully, or somewhere in between the two. We elected to down weight those observations identified by raising the weights to the 0.707 power (square root of one-half). This allows those observations to have an impact on the recalculation of the group design effect but their effect will be much smaller than before down weighting.

There were three primary items on which we compared these five different methods. The first item was how the new methods compared to the procedure in 1990. The second item was how the weighted average and the median of the RAD values compared between the methods. The last item is that of aesthetics.

In comparing the results to the 1990 methodology (or at least part of it), Q_n agrees most closely on identifying observations with RAD values which would have been candidates for outliers using the 1990 method. Both MAD_n and S_n frequently do not identify observations as outliers which have RAD values of 80 percent or more. This is a real issue since that means the published SE will disagree with the actual SE by that same amount. So on this comparison, Q_n is better than MAD_n or S_n for our data. It does not make sense to compare the influential methodology to the 1990 results in this setting.

Q_n had the lowest weighted average of the RAD values and the lowest median RAD values of all the methods used including the influential observation methodology. This suggests again that Q_n is the best methodology using our criteria.

Aesthetically, the influential observation methodology is the best because of the continuous down weighting scheme it employs. With the 1990 method, MAD_n, Q_n, or S_n, there is always the issue of how to choose the cutoff for determining outliers. Since the influential observation methodology applies a smooth down weighting there is no magical cutoff value. This removes the issue of one observation just below the cutoff being treated differently than an observation which is just higher.

7 Conclusions

For our situation, it appears that S_n and Q_n are the most effective for detecting extreme values given the asymmetric nature of the data. They also performed the best in comparison to the 1990 method by more consistently identifying observations with high RAD values. Their performance did vary from site to site, however.

One result of our research was that the most aesthetic solution did not give the most pragmatic solutions. The influential observation methodology appeared to give a very reasonable approach but in comparing median and weighted average RAD values for groups it did poorer than both S_n and Q_n. One possible explanation for this is that this method might have a tendency to reinforce the original group design effect. This makes the down weighting less effective and the result is a less representative recalculated group design effect.

We plan at this time to test all five methods again on the Census 2000 Dress Rehearsal data when it is available. Because the Dress Rehearsal data is collected and tabulated slightly differently, it is possible that a different methodology may be more suited to the long form data than the ACS data. These results will then be used to determine which method is used for the Census 2000 operation.

7.1 Items for further research

In this paper we presented just a few ways to compare the different methodologies in identifying extreme values or influential observations. More work needs to be done on this to not only allow better comparison of methodologies but also to provide tools for assessing the quality of the implementation of our outlier methodology in production. A number of graphical comparisons have been made but those must be reviewed to determine which graphs truly help and which do not.

Another area for research is using other estimators of location to determine the group design effect. We know that the mean and hence the weighted mean are easily influenced by extreme/influential values and thus it may be worthwhile to investigate more robust estimators of location.

Finally, whatever additional methods we may find for comparing the quality of our outlier methodology needs to be quantifiable so that an objective means can be used to compare both across methodologies and within methodologies. This allows us to not only decide on the best method but also on how to best employ that method.

References

Belin, T. A., Schenker, N., and Zaslavsky, A. M. (1999), "Down Weighting Influential Clusters in Surveys, with Applications to the 1990 Post-Enumeration Survey." Draft manuscript.

Croux, C., and Rousseeuw, P. J. (1992), "Time-efficient algorithms for two highly robust estimators of scale," in *Computational Statistics, Volume 1*, eds. Y. Dodge and J. Whittaker, Heidelberg: Physica-Verlag, 411–428.

Huber, P. J. (1981), Robust Statistics, New York: Wiley.

Rousseeuw, P. J., and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *JASA*, 88, 1273–1283.