TRIMMING EXTREME WEIGHTS IN HOUSEHOLD SURVEYS

Benmei Liu, David Ferraro, Erin Wilson, and J. Michael Brick, Westat
Benmei Liu, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Outliers, influential observations, household surveys, detection methods

1. Introduction

An outlier in sample surveys occurs when an extreme value is observed for a characteristic, the weight for a sampled unit is very large relative to the weights of other units, or there is a combination of a relatively unusual observed value and a large weight. To be considered an outlier, the case should be influential—in the sense that it has a major effect on analysis of the data. However, the influence might be in estimating the variance of the estimate and, thus, affecting inferences. If an observed value is unusual and the sampled unit has a relatively small weight, then the influence of the case on most analyses may not be sufficient to classify it as an outlier.

Outliers may appear even if the sample design, data collection, and data preparation are carefully crafted and implemented. As noted by Lee (1995), outliers in sample surveys may be either representative or nonrepresentative. Representative outliers are those correctly recorded and represent other population units similar in value to the observed outliers. Nonrepresentative outliers are those that are incorrectly recorded or unique in the sense that there are no other units like them. We consider only representative outliers. In addition, it is worth noting that observations that are not outliers when full population estimates are produced may be outliers for estimates of domains. This further complicates the problem of identifying outliers.

This paper focuses on methods of identifying cases in household surveys as outliers because they have large weights, rather than dealing with unusual observed values of characteristics. There are several reasons for this focus. First, large weights are most likely to have a substantial influence for a variety of analyses, especially for estimates of domains. Second, at the time of data processing, it is possible to identify large weights and trim or otherwise deal with them. In some circumstances this may not be possible with the observed characteristics. Third, there are many existing procedures that recommend ways of identifying and dealing with observed values that are unusual. These procedures generally do not deal with unequal weights. Thus, there is a greater need to deal with the identification of influential weights in a sampling setting.

Outliers due to extremely large weights most often occur when the design samples units with different probabilities to meet some design goal. For example, a survey may sample household members with different probabilities to achieve specified sample size goals by age, or adults may be sampled with different probabilities across households so that no more than one adult is selected within a household. If outliers can be identified during the weighting stage of the survey, it is possible to make an adjustment such as trimming the weight that will eliminate some problems during subsequent analysis of the data.

Potter (1988, 1990) and Lee (1995) review and propose methods for dealing with outliers in sample surveys, where the outliers may be due to unusual observed values or large sampling weights. Despite the suggestions they and others have given, most surveys still use *ad hoc* procedures that may be arbitrary and not supported by any theory in order to identify outliers. One of the reasons for this may be that the methods suggested in the literature are not fully defined in the sense that they do not provide specific criteria for classifying observations as outliers. Consequently, the evaluation of alternative methods is very limited.

The next section examines methods of identifying outliers in household sample surveys. We begin by reviewing some of the methods suggested in the literature and discussing some of their limitations. New methods are then proposed that are more relevant to identifying cases that have large and influential weights. The methods rely heavily on the ranking of the weights. We propose specific procedures to identify outliers that may be portable across different household survey designs and may eliminate some of the arbitrariness that is often associated with this task. Thus, guidelines for classifying units with large weights as outliers are proposed. The third section implements the proposed methods using two complex surveys conducted by Westat. The strengths and the weaknesses of the methods are evaluated. The last section gives some conclusions and suggestions for additional research.

2. Outlier Detection Methods

2.1 Existing Methods from Literature Review

The literature for identifying and modifying outliers in household survey data is relatively limited. Most methods have been designed to handle survey specific situations and can not be successfully applied

to general household survey designs. Some of the methods are discussed in this section, as well as new methods devised throughout the course of this research. Some methods could be used to identify very large weights **and** very small weights. As discussed in the introduction, only large weights are considered in this study because small weights do not have as much of an effect on sample estimates.

An extreme weight may be declared as an outlier based on its relative distance from the center of the data. For instance, let $distance_i = \frac{|w_i - m|}{s}$, where w_i is the weight for sampled unit i, m is a location measure representing the center of the weights, such as the median, and s is a scale measure. One common candidate for the scale measure is the median absolute deviation, defined as

$$AD = median\left\{\left|w_i - \underset{j}{median}\left(w_j\right)\right|\right\}.$$

When an observation has a large value of $distance_i$, it indicates that the weight is relatively large in comparison to the other weights in the dataset.

Another method for identifying outliers is the forward search method described by Chambers (2003). Assuming a survey data set of size n, the algorithm begins with an initial subset of $m<n$ observations which are assumed to be "clean", meaning not containing any outliers. Using this subset, a regression model for the variable of interest is fit. Deviations of the fitted values generated by the model from the actual sample values are then calculated, and a new "clean" subset is formed containing the observations that produce the m+1 smallest distances. This procedure is repeated until the calculated deviations of the observations outside of the clean subset are all considered to be too large, and therefore outliers, or until the subset is exhaustive.

Potter (1990) describes an outlier identification method that does not use actual survey data, but rather relies on an assumed distribution of the weights. In this procedure, a trimming level is pre-specified based on a probability of occurrence according to the distribution model. For instance, if the trimming level is set to 1 percent, any observation whose probability according to the model is at most 0.01 will be considered an outlier and a candidate for trimming. Another option is to implement this procedure for an initial trimming and then distribute the excess weights among the untrimmed cases. The parameters of a new sampling weight distribution can then be estimated and a revised trimming level set. A second trimming can then take place followed by another redistribution of the excess weights to untrimmed cases.

Another popular type of outlier detection procedure involves examining the contribution of each individual sampling weight to the sum measure of entropy. Those weights that contribute substantially to the entropy are considered outliers. One such method, referred to by Potter as the NAEP procedure, identifies outliers based on the contribution to the overall sampling variance. This is accomplished by computing a value using the sum of the squared weights, $c\sum w_k^2 / n$, where c is a preset constant and n is the number of observations. For each observation, the squared weight is compared to the above quantity and those cases exceeding that trimming level are trimmed to the square root of that value. The excess weight is redistributed to untrimmed cases to retain the sum of the weights. The process is then repeated until no case remains with a squared weight that is larger than the trimming level. A similar method compares each sampling weight to some value, k, times the median of the sampling weights. The median is used instead of the mean because the mean can be heavily influenced by extreme weights. Often, k is set to be a simple constant such as 3 or 5, but it can also be defined by the distribution of $\{w_i / median,\ i = 1, 2, \ldots, n\}$. The weights larger than the trimming level are trimmed to that value and the excess weights are redistributed among the untrimmed cases. To limit the number of cases to be trimmed, k can be increased.

The methods described above are the primary ones that we considered based on the literature review, although there are certainly several others available. Our review suggested that none of these methods as described in the literature could be automated and consistently identify observations that would be good candidates for trimming. Most of these procedures were not implemented in our study because some feature did not make them useful for this study, or because they did not apply well to household survey data. Several of the methods were designed to be more useful for identifying outliers in characteristics rather than outliers in survey weights. For example, the forward search method described by Chambers was designed to identify extreme observations (y-values), not extreme weights. An attempt was made to modify the procedure to handle the weights; however, even after several iterations, the largest deviations did not always correspond with the largest weights. Therefore, the "clean" subset could not be relied upon to include all nonoutlying cases. The weight distribution procedure does specifically address outliers in weights, yet this procedure also was problematic. The method requires that the distribution of the weights be either known or accurately estimated. A few known distribution models including the ones suggested by Potter were tested to

see if the household survey data in our work fit these models. However, no reasonable fit was found, and the method was not explored any further. With both the NAEP procedure and the *k*median* method, a constant is required in order to determine the trimming level. We examined some constants, but found that household survey weights vary considerably across subgroups. No preset value seemed appropriate across household survey designs and this limits that ability to automate the procedure and make it portable.

2.2 New Techniques

In addition to the methods found in the literature, other methods that were not specifically designed for sample surveys were reviewed and adopted for this purpose. The procedures that were deemed to be most appropriate were those that use the spacings between the weights as a means of identifying outliers. See David (1970). To implement these procedures, the weights are first ranked from largest to smallest. Using order statistic notation (i.e. $w_{(n)}$, where n is the number of observations), the four largest weights in the dataset from largest downward are: $w_{(n)}$, $w_{(n-1)}$, $w_{(n-2)}$, and $w_{(n-3)}$. A "spacing" is the distance between a ranked weight w_i and the next ranked weight $w_{(i-1)}$, i.e., the spacing $z_{(i)} = w_{(i)} - w_{(i-1)}$.

Two new methods for identifying largest weights in household surveys were developed using this concept of spacings. The first of these methods aspires to identify large gaps in the weight distribution for the largest of the weights. For each weight, the spacing between it and the next largest weight is compared to the spacings between the next five pairs of ranked weights. The value

$$d5\_space_{(i)} = \frac{z_{(i)}}{z_{(i-1)} + z_{(i-2)} + z_{(i-3)} + z_{I-4)} + z_{(i-5)}}$$

increases when an observation is considerably larger than the next largest weight, in comparison to how much the next few weights vary from each other. This measure shows when there are large jumps in the distribution of the weights, which is an indicator that the weight is an outlier and should be considered for trimming. The second spacings method examines the distance between a weight and the next largest weight relative to the size of the weight. After some examination of an appropriate measure, we defined $rel\_space_{(i)} = \frac{z_{(i)}}{w_{(i)}} \times 10$. This definition allows for the procedure to be implemented in the same way for different groups, regardless of how the magnitude of the weights may differ across subgroups.

Another method that is closely related to the NAEP procedure described by Potter is proposed to measure the effect on the variance estimates, by examining the effect of dropping a particular weight. This method, called the RV method, compares an estimate of the effective sample size, as a function of the relative variance, given that the ith weight is dropped. After several iterations, we decided on the formulation given below:

$$RV_{(i)} = \frac{\hat{Effss}_{(i)} - \hat{Effss}_{(i-1)}}{\hat{Effss}_{(i-1)} - \hat{Effss}_{(i-2)}},$$

where

$$\hat{Effss}_{(i)} = \frac{i}{1 + rel\_\mathrm{var}_{(i)}}.$$

For example, when calculating $RV_{(n)}$ of the largest weight $W_{(n)}$, $\hat{Effss}_{(n)}$ will be calculated using all observations. The quantity $\hat{Effss}_{(n-1)}$ will be calculated using n-1 observations after dropping the nth or largest weight. To calculate $\hat{Effss}_{(n-2)}$, n-2 observations are used, after dropping the n-1 and nth observations, or the two largest weights.

2.3 Composite Score

The new methods described above and the methods from the literature search were explored using a household survey dataset. No single measure was found that clearly identified the vast majority of cases that were deemed outliers without identifying far too many cases that should not have been classified as outliers. The strategy followed was to develop criteria for trimming based on a combination of the methods that each was deemed useful for household survey data. During this process, a variation of the relative distance method was implemented. The measure is the spacing between the relative distance for a weight and the relative distance for the next largest weight. That is, $diff\_dist_{(i)} = distance_{(i)} - distance_{(i-1)}$, where $distance_{(i)}$ was defined earlier as the relative distance for weight $w_{(i)}$.

Several different criteria were developed and tested, involving the three new measures (diff_dist, d5_space and rel_space). There was not enough consistency in the behavior of RV to support including it as part of the criterion. Instead, it was treated as a

source of additional information to help make decisions about cases that were questionable candidates for trimming.

The final procedure identified any observation meeting **all** of the following criteria as a candidate for trimming:

- diff_dist ≥ 1.0;
- d5_space ≥ 0.9; and
- rel_space ≥ 1.0.

For the observations meeting all three criteria listed above, a composite score was calculated. The composite score is the sum of the values of each of the three measures (diff_dist, d5_space and rel_space). The score also includes a "penalty" that reduces the score as the number of cases to be trimmed increases. The rationale for the penalty is as follows. When any observation with a score above a certain level is considered for trimming, it also implies trimming all of the weights that are ranked higher. In order to reduce the chances of trimming too many cases, a penalty of $\frac{n - \text{rank}}{2}$ (where rank is n, n-1, n-2, …, 1, for the weights from largest to smallest) is deducted from the initial score. In this way, there is little to no penalty for trimming a few cases and a larger penalty for trimming more.

After evaluating different scores based on data from two surveys discussed in the next section, a scale was devised to aid in making trimming decisions. Even though the goal was to develop a fully automated procedure, it became evident that in many situations, some additional scrutiny may still be required to make a decision for the questionable cases. Thus, three levels associated with the scale score were proposed. They are:

- **Score > 8** – Automatic – these cases are considered definite outliers that should be trimmed
- **Score between 4 and 8** – Questionable – these cases have extreme weights by at least some of the criteria, but before the decision to trim is made, further evaluation is needed. The RV measure may be useful along with visual review of graphs of weights.
- **Score < 4** – No Action – these cases generally should not be trimmed.

The scale and proposed cut-offs are examined in the next section.

3. Empirical Study

3.1 Study Design

To evaluate the proposed outlier detection methods for household surveys, we use data from two random digit dial (RDD) telephone surveys. The first survey is the National Survey of America's Families (NSAF) conducted by Westat for the Urban Institute to study status of families as changes in social programs were implemented beginning in the late 1990's. The survey collected information on the economic, health, and social dimensions of the well-being of children, adults under the age of 65, and their families in 13 states and the balance of the nation. There were three rounds of data collection: 1997, 1999 and 2002. For more information on NSAF, see the Urban Institute website listed in the references that contains a variety of methodological reports on the survey design and weighting. The second survey is the California Health Interview Survey (CHIS), a collaborative project of the UCLA Center for Health Policy Research, the California Department of Health Services, and the Public Health Institute. In this survey Westat telephone interviewers collected information on if, where, and how people get health care in California. The sample was allocated by county and aggregates of smaller counties with supplemental samples of selected populations and cities to form 41 sampling strata. There have been two rounds of data collection: 2001 and 2003. For more information on CHIS, see the UCLA website listed in the references that contains methodological reports for both 2001 and 2003 surveys.

These two surveys were chosen for several reasons. First, the surveys exhibit differential selection probabilities that vary by strata or subgroups. Second, both surveys have the features of multiple weights (for different groups such as adults and children). Third, both surveys have been conducted more than one time. These features allow us more opportunity to evaluate the proposed methods under varying circumstances and, thus, better assess the portability of the methods.

As noted earlier, outliers due to extremely large weights in household surveys are typically due to large unequal probabilities of selection. These surveys also have multiple weighting adjustment factors, such as nonresponse and poststratification, but the sizes of these adjustments are usually controlled more by various choices in the weighting, such as the selection of nonresponse adjustment cells. The weights for the NSAF and CHIS are typical household surveys in this respect. To illustrate the variability in the probability of selection weights, consider sampling children in CHIS. In this survey if a household has children under age 12, one child is selected for the sample. Thus, if a household has 7 children under age 12, the weighting

factor for the selected child is 7, the inverse of the probability of selection. On the other hand, if a household has only one child under 12, then that child would have a factor of 1.

Differential sampling weights in CHIS are associated with the following selection probabilities:

- Initial sampling by strata;
- Oversampling by ethnic population; and
- Sampling person within household.

Similar features are present in NSAF, where the probabilities of selection include differential factors associated with:

- The initial sampling by state;
- Subsampling by income level;
- Subsampling households without children; and
- Sampling person within household.

3.2 Evaluation

In this section we describe the findings of the empirical study using data from NSAF and CHIS. As alluded to in the earlier sections, the development of the methods of detecting outliers was exploratory in the sense that several approaches and measures were considered and evaluated prior to deciding on the specific ones. Using the same datasets to determine the procedures and to evaluate them is a bit unusual. To ameliorate this problem, the exploratory work was done on a dataset from one survey, then evaluated using a dataset from the other survey. Thus, the methods are probably more robust than might have otherwise been true. Nevertheless, we consider all the developments to be exploratory and there is a definite need for more rigorous evaluation.

One of the key approaches to the study was to treat major subgroups from each of the surveys separately. In the NSAF, 13 geographic areas have probabilities of selection that vary widely, and these areas are treated as separate subgroups in the analysis. In the CHIS, each of the 41 geographic areas (counties or groups of counties) are treated separately because the rates differ greatly across these strata. While this is a survey-specific aspect of the application, at least this level of adjusting to the specifics of the survey may be necessary to gain some portability across surveys.

Before presenting the findings, it may be useful to describe some of the motivation for the approaches we developed. The main motivation was our experience in reviewing graphical outputs to identify outliers from sample surveys. Invariably, graphs of the weights in a survey are revealing, and the gaps or spacings between the largest weights is a natural criterion considered in visually identifying outliers. If the gaps are large, then trimming may be needed.

One typical structure is that the largest weight is very distinct in the graph and appears as an obvious outlier. This type of situation can be easily detected using most detection methods including the three criteria and scale score method we propose. However, even in this case it is not always easy to determine if the largest weight is so large that it should actually be trimmed. Consequently, in applied work the treatment of the weights may not be consistent. When there are several large weights in the tail of the weight distribution – a situation that entails more complexity than the single large weight case – the size of the spacings of the weights are very influential in determining which weights should be treated as outliers. In both cases, the development of automatic detection methods is useful.

Because of space limitations in this paper, we limit the graphs and tables presented here. One graph and two tables from the research were selected to illustrate several of the key concepts. The first graph is from the 1999 child survey in the Balance of U.S. (the set of states that were not identified as a separate geographic sampling area) for the NSAF. A table from the same source is also presented. In addition, data from the 2003 CHIS adult survey for selected geographic strata are presented in a table.

Figure 1 shows the three measures for the thirteen largest weights in the Balance of U.S. sample from NSAF. Notice that the measures jump up for the fifth largest weight (n-4). This weight satisfies all the three criteria listed in the previous section and, therefore, constitutes a potential cut point. This graph also demonstrates that it is possible for more than one weight to satisfy all the three criteria, in this case the largest and 5th largest weight satisfy all three conditions. This implies that it is necessary to examine a relatively large number (say 25) of individual weights to find the last one in the series that meets all the three criteria for being a cut point. Using the three proposed criteria, the 5 largest weights are identified as potential outliers and considered candidates for trimming.

Once the candidate weights for trimming are identified, the scores for those that meet all three criteria are computed. Table 1 gives the scores for the 7 largest weights in the sample for the Balance of U.S. along with some other details not shown in Figure 1. The diff_dist for the largest weight is 9.2, which is very

large compared to the cut-off score, whereas the measure of d5_space is only 0.9. This example highlights why it is important to look at the three measures together to identify outliers consistently.

Figure 1 and the numbers in Table 1 reveal why the three measures are so different for the largest weight in this sample. The value of diff_dist depends on the distribution of all the weights in the sample, while the measure of d5_space depends entirely on the 6 adjacent weights. As shown in the table, there are a number of sizeable gaps among the largest weights, which is why the d5_space is relatively small in this case. The score for the largest weight and for the fifth largest weight (12.4 >8) is in the automatic trim range. Thus, the 5 largest weights should be trimmed according to the proposed guidelines.

The table also contains some other statistics that help explain the variation in the weights. The variability of the weights is largely due to the sampling of one child from a household where some households have a large number of eligible children. The RV of the weights shown in the table are consistent with the three main criteria in this example. The RV for the largest and 5th largest weight are much greater than for the other weights. In other situations we found that this was not always the case and the RV could provide some independent information that could be used to help determine whether to trim weights that had scores in the range of 4 to 8. Despite our belief that the RV is valuable, we have not been able to develop a reasonable criterion for using this measure directly in identifying outliers.

Table 2 shows summary data similar to that in Table 1, but these data are from select strata from the 2003 CHIS adult file. The strata were chosen to include many of the different situations that were encountered in this study. Some strata had no weights that met all 3 criteria so no candidate outliers were detected. We have not included any of these strata in the tables. Below we review the issues in each stratum.

In Stratum 3 data are given for the six largest weights. Only the 2 largest weights meet the three criteria and the scores for both of them are greater than 8. Following the guidelines, these two points should be trimmed. In Stratum 4, the 10 largest weights are shown because both the largest weight and the 9th largest weight (rank=*n*-8) satisfies all three criteria simultaneously. The 9 largest weights are candidate outliers, but the score for the 9th weight is just negative so only the largest weight should be trimmed because it is the only one with a score of greater than 8. The next stratum, Stratum 7, has 11 observations listed in the table. In this situation the first weight that meets all three conditions is the 10th largest weight, and the score for that weight is small (<4). Thus, no weight should be trimmed according to the guidelines. Note that in this stratum, the penalty had the effect of moving this point from the questionable category to the no action category. In Stratum 11 the largest weight is the only candidate outlier according to the three criteria, and it has a score of 6.6 that is in the questionable range. The RV criterion may be useful to inform the trimming decision. When we carefully examined this weight, we came to different conclusions about the need for trimming and believe that this may be contingent on the uses of the survey. The questionable range seemed appropriate for this weight. Stratum 15 shows a situation in which the largest weight should be definitely trimmed, and the second largest weight falls into the questionable range. In this stratum either one or two weights might be trimmed.

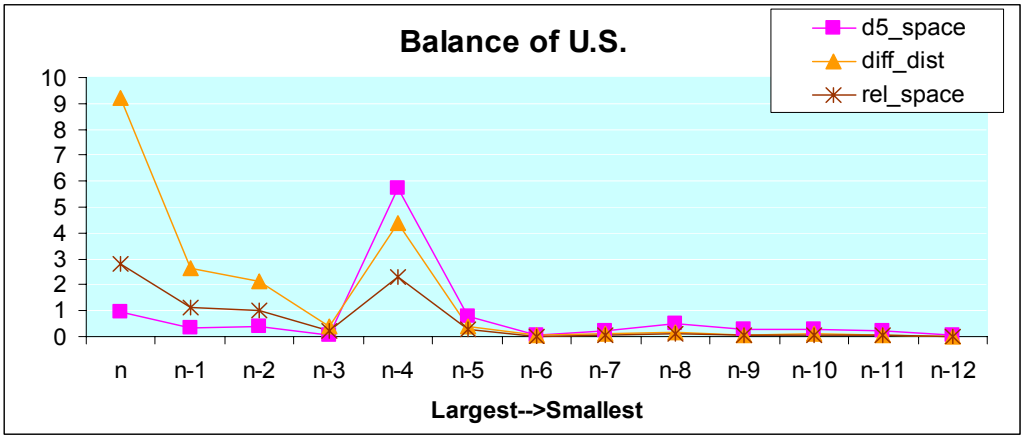

Figure 1. The three measures in the Balance of U.S. stratum from the 1999 child NSAF file

Table 1. Statistics from the Balance of U.S. stratum from the 1999 child NSAF file

Rank	Weight	# of kids	diff_dist	d5_space	rel_space	RV	Score
n	63,109	6	9.2	0.9	2.8	2.2	12.9
n-1	45,528	7	2.7	0.4	1.1	1.4	--
n-2	40,470	8	2.1	0.4	1.0	1.3	--
n-3	36,423	6	0.4	0.1	0.2	1.1	--
n-4	35,696	7	4.4	5.7	2.3	2.4	12.4
n-5	27,381	2	0.4	0.8	0.3	1.1	--
n-6	26,653	5	0.0	0.1	0.0	1.0	--

Note: Score is only computed for weights that satisfy all three criteria.

Table 2. Statistics from selected strata from the 2003 CHIS adult file

Stratum	Rank	Weight	diff_dist	d5_space	rel_space	RV	Score
3	n	17,163	19.5	2.0	4.2	3.6	25.7
3	n-1	9,877	7.2	2.7	2.7	2.3	12.1
3	n-2	7,198	0.7	0.3	0.4	1.1	--
3	n-3	6,940	0.2	0.1	0.1	1.0	--
3	n-4	6,871	1.4	0.6	0.8	1.2	--
3	n-5	6,339	0.4	0.2	0.2	1.1	--
4	n	6,970	4.9	1.2	2.0	1.8	8.1
4	n-1	5,542	1.3	0.3	0.7	1.2	--
4	n-2	5,167	0.7	0.2	0.4	1.1	--
4	n-3	4,974	1.7	0.4	1.0	1.3	--
4	n-4	4,484	0.0	0.0	0.0	1.0	--
4	n-5	4,478	0.5	0.2	0.3	1.1	--
4	n-6	4,321	1.7	0.7	1.2	1.4	--
4	n-7	3,821	0.2	0.1	0.2	1.0	--
4	n-8	3,763	1.4	1.1	1.1	1.4	-0.4
4	n-9	3,364	0.0	0.0	0.0	1.0	--
7	n	4,740	4.8	0.4	0.8	1.2	--
7	n-1	4,380	3.5	0.2	0.6	1.1	--
7	n-2	4,118	0.1	0.0	0.0	1.0	--
7	n-3	4,113	6.7	0.6	1.2	1.4	--
7	n-4	3,613	0.1	0.0	0.0	1.0	--
7	n-5	3,608	3.3	0.3	0.7	1.2	--
7	n-6	3,360	4.5	0.5	1.0	1.3	--
7	n-7	3,020	0.4	0.0	1.1	1.0	--
7	n-8	2,992	2.0	0.3	0.5	1.1	--
7	n-9	2,842	4.0	1.5	1.1	1.4	2.1
7	n-10	2,539	0.6	0.2	0.2	1.0	--
11	n	3,799	3.0	1.7	1.9	1.9	6.6
11	n-1	3,068	0.1	0.1	0.1	1.0	--
11	n-2	3,033	0.7	0.4	0.5	1.2	--
15	n	6,749	2.3	0.7	5.4	1.9	8.4
15	n-1	5,213	2.1	0.9	3.8	1.8	6.3
15	n-2	4,136	0.2	0.1	0.2	1.0	--

Note: Score is only computed for weights that satisfy all three criteria.

4. Conclusion

Our goal was to develop a method for automatically detecting outliers due to large survey weights that should be trimmed for applications to household surveys. In reviewing the literature, we noted that many of the outlier detection methods were constructed more for detecting unusual values of characteristics than for dealing with weights. The methods that did deal with the weights directly tended to be survey specific and required a fair bit of customization to the particular survey situation. As a result we examined some other alternatives, based largely on spacings of the weights. When we examined these methods, we observed that they had difficulty identifying outliers without including a number of observations that might not be outliers.

As a result, we proposed a composite method. First, three criteria based on different outlier detection procedures were established and candidate outliers are those cases with weights that meet all three criteria simultaneously. A summative score with a penalty that increases with the number of identified outliers is then computed for the candidate outliers. All the observations with scores of greater than 8 are considered outliers that should be trimmed. Those with weights of less than 4 should not be trimmed. Those weights with scores between 4 and 8 are in the questionable range and survey specific goals or other measures such as RV may guide the trimming decision for these points.

This method does not satisfy our initial goal of having a fully automated system, but we believe it greatly reduces the burden for outlier detection and may be portable across many household surveys. Further work needs to be done to evaluate the proposed method in other surveys and perhaps refine it. An obvious need is revision of the RV procedure or the development of another measure that can address the effect of the outliers on the variances of the estimates. Another clear need is to better place these practical methods into a theoretical framework.

After the outliers are identified by methods such as those proposed here, the weights of the outliers are trimmed. One issue that we did not address is what to do with the other weights when the largest weights are trimmed. If the other weights are not adjusted, then the sum of the weights of the survey will be biased downward. In many household surveys, trimming is the last weighting adjustment before poststratification. If this is the case, then an option is to poststratify the trimmed weight to known population control totals. The excess weight which was trimmed is redistributed as part of the poststratification. If weight trimming is an intermediate step in the weighting process, then the method of dealing with the excess weight may be more difficult. The trimmed weights may or may not be redistributed to preserve the weighted totals, depending on the specifics of the survey.

5. References

David, H. (1970). *Order Statistics*. John Wiley & Sons, New York.

Lee, H. (1995). Outliers in Business Survey, *Business Survey Methods*, John Wiley & Sons, New York.

Potter, F. (1988). Survey of Procedures to Control Extreme Sampling Weights, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 453-458.

Potter, F. (1990). A study of Procedures to Identify and Trim Extreme Sampling Weights, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 225-230.

Chambers, R, Hentges, A., and Zhao, X. (2003). Robust Automatic Methods for Outlier and Error Detection, *Journal of the Royal Statistical Society*, A, 167, 323-339.

6. Websites

The California Health Interview Survey (CHIS): www.chis.ucla.edu

The National Survey of America's Families (NSAF): www.urban.org/Content/Research/NewFederalism/NSAF/Overview/NSAFOverview.htm