Treatment of Influential Values in the Annual Survey of Public Employment and Payroll

Joseph Barth, John Tillinghast, and Mary H. Mulry[1]
U.S. Census Bureau
joseph.j.barth@census.gov

Abstract

Like most surveys, influential values occur rarely in the U.S. Census Bureau's surveys of governments. However, when they do occur, they have a large impact on survey estimates. In order to improve the quality of estimates for the Annual Survey of Public Employment and Payroll (ASPEP) we consider two robust survey estimation methods: M-estimation, and Clark Winsorization. These methods are compared to the Generalized Regression method and the classic Horvitz-Thompson estimator. Since the Census of Governments: Employment Component (COG-E) collects the same variables collected by the ASPEP, our evaluation uses a simulation approach that draws repeated samples from the intersection between the 2002 and 2007 COG-E universes following the new 2009 ASPEP sample design. We estimate mean squared error, standard error, and bias for each estimator, as well as the estimated values themselves, as measures of comparison for the various methods.

1. Introduction

Influential values occur infrequently in the U.S. Census Bureau surveys of governments but are problematic when they do occur. An observation is considered influential if its value is correct but its weighted contribution has an excessive effect on the estimate of interest. These observations tend to have a large weight when compared to other cases. A recent redesign of the Annual Survey of Public Employment and Payroll (ASPEP) may have increased its vulnerability to influential values even though the chance of occurrence remains small. This is the motivation for investigating methods for detecting and treating influential values in the ASPEP.

The ASPEP provides current estimates for full-time and part-time state and local government employment and payroll by government function (e.g. elementary and secondary education, higher education, police protection, fire protection, financial administration, etc.) This survey covers all state and local governments in the United States; this includes counties, cities, townships, special districts, and school districts. The first three types of governments are referred to as general-purpose governments as they generally cover several governmental functions. School districts cover only educational functions while special districts usually cover one, but sometimes two or more functions (e.g., sewer and water). Data on employment include number of full-time and part-time employees, gross pay, and hours paid for part-time employees. For each government, data are reported for the pay period that includes March 12.

In 2009 the ASPEP sample underwent a significant redesign in order to reduce the number of small governments included in the sample. While the results of this redesign were largely positive, one drawback was an increase in the weights of small sample units. This change increased the survey's sensitivity to influential values; small increases in the employment levels of small governments now have a greater chance of creating significant overestimation of state totals.

Research in the last few years regarding influential values in U.S. Census Bureau's Monthly Retail Trade Survey (MRTS) (Mulry and Feldpausch 2007a, 2007b) showed that two methodologies, M-estimation (Beaumont and Alavi 2004) and Clark Winsorization (Clark 1995), had potential for improvements in systematic detection and treatment of influential values in the MRTS. Further research (Mulry, Oliver, Kaputa 2011; Mulry and Oliver 2009) is applying the two recommended methods to simulated data to examine the statistical properties of the treated estimates obtained using each method over repeated sampling.

[1] This report is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

In this paper, we compare two methods for treating influential values, Clark Winsorization and a variation on the application of M-Estimation, against non-robust expansion and generalized regression (GREG) estimation. These methods are applied to repeated samples drawn from the 2007 Census of Governments: Employment Component (CoG-E) and conclusions are drawn from the empirical distributions of quality statistics generated by these samples.

2. Background

2.1 Survey and Census Background

The CoG-E is a census conducted every five years that collects the same data as the ASPEP would for that year. The ASPEP is an annual survey of all state and local governments in the 50 states, plus Washington, D.C. The universe and frame are the same as those used in the CoG-E, with updates made to reflect any births, deaths, or mergers that may have occurred. A unit is determined to be a government if it exists as an organized entity, has governmental character (such as the power to levy taxes), and displays substantial autonomy (i.e., considerable fiscal and administrative independence).

The ASPEP and CoG-E both collect data on five variables, and derive two additional variables from these. The five variables collected are full-time employees, full-time pay, part-time employees, part-time pay, and hours worked by part-time employees. The first derived variable is total pay, which is simply the sum of full-time and part-time pay. The second is full-time equivalent, which is calculated by dividing the number of part-time hours worked by the standard number of hours in a workweek for full-time employees in the particular government, added to the number of full-time employees in that government.

The data for each unit are subdivided into twenty-three different items, such as fire protection, sewerage, and hospitals. Not every unit has all twenty-three items. For instance special districts and school districts typically only have one or two items.

ASPEP estimates are published at national and state levels for state-only, local-only, and state-and-local aggregates. For example, we can view estimates for just state governments for Alabama, estimates for just local governments for Alabama, or all state governments combined with all local governments in Alabama. We can view an estimated national total for all state governments, an estimated national total for all local governments, or an estimated national total for all state and local governments combined. If we do not consider data from Washington, D.C. this gives us 150 state level estimate tables, and 3 national level estimate tables.

2.2 Data for Analysis

In our analysis we focus on the number of full time employees for the aggregated total over all functions. The sum of full-time pay and part-time pay from a prior year are used as a measure of size for the unit. We use data from the CoG-E to create simulated samples using real data.

The universe for our study is the intersection of the 2002 and 2007 CoG-E universes. We use the data from the 2002 CoG-E portion of this intersection to create our sampling frame. We stratify units by state and type of government, and then further stratify into certainty and sample substrata, with large units being placed in the certainty stratum. We then sample within sample substrata using inclusion probability proportional to size (PS) sampling with total pay, the sum of full- and part-time pay, as our measure of size. Units with a measure of size equal to 0 are sampled via simple random sampling without replacement. We take repeated samples from this scheme and base our analysis on a comparison of sample values to the true value, and the empirical distribution of various measures of quality.

3. Methodology

3.1 Winsorization

Winsorization, at its most basic, is simply replacing any value that is above a specified threshold with a value equal to the threshold. This threshold could be determined by a logical argument, subject matter expertise, or statistical analysis. The threshold could be the same for all units or could be different for separate subgroups of the population (even down to individual units.) We consider individual unit thresholds that are designed to minimize the mean square error (MSE) of the winsorized estimators under a very general model using a method developed by Clark (1995).

We now wish to state the result of Clark that we use for our winsorization. Note that the optimal thresholds depend on the bias of the estimator, which in turn depends on the thresholds used. We use an iterative algorithm that we will describe shortly.

Let , where denotes the sample, be a linear unbiased estimator of the population total *Y* under the model:

Let . Define for some fixed ; define the winsorized estimator as:

Then the value of which minimizes the MSE of are such that:

Where

Following Clark's suggestion we use an approximation to the optimal

We calculate these approximate optimal cutoffs by using the method in Chambers, et.al. (2001), which was inspired by Kokic and Bell (1994). Define with some estimator of and the values of sorted in descending order. Then define

We then define our approximate optimal cutoff threshold estimator to be

It can be shown that is in fact an estimator of the bias due to winsorization under our assumed model; see Chambers, et.al. (2001) for details.

We look at three possible estimators of ; the simple stratum mean; the stratum median; and a simple robust regression estimate based on the ratio model,

Where denotes the subset of sampled units consisting of those units which have a non-zero value for full-time employees in both the 2002 and 2007 CoG-E. is the value of the survey unit in the 2002 CoG-E, and is defined similarly. This is done to remove a large number of zero-valued ratios which often cause problems, as well as the obvious necessity of dealing with division by zero.

3.2 Generalized Regression (GREG)

Because we do a Census of Governments every five years, there is auxiliary information available for almost all units—the employment totals from the last CoG. Having auxiliary information for everyone means that we can use regression-based approaches, of which GREG is one of the simplest (Sarndal, 1992). We simply model the current survey year's data in terms of the CoG-E year's data, regress on the sample data, and use the resulting coefficients to estimate current totals.

Formally, this means that we estimate and by minimizing the objective function

where are the design-based weights. Then our estimated total is , with the standard expansion estimator of .

3.3 Robust GREG (M-estimation)

This method is similar to GREG, but it minimizes a more robust objective function so that influential points and outliers do not have such a disproportionate effect on the estimate. Because the objective function is not a pure quadratic, we use an iterative method to find the minimum.

The function is of the form

Where

are again the design-based weights;

is a spline function, quadratic for values small in absolute value and linear for values larger in absolute value than some positive number K, typically (see Figure);

and is a scaling parameter determined iteratively.

The solution is found by iterative reweighted least squares (IRLS). For initial ( ) we use the result of the GREG estimator.

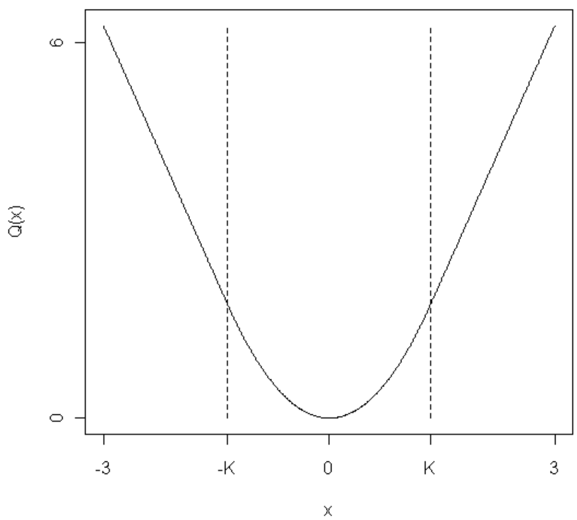


Figure. The robust objective function Q is quadratic up to abs(x)=K; then it extends linearly. Q' is continuous; Q'' is not.

Minimizing is equivalent to setting these two sums to zero:

These are the same conditions that minimize

where and are weight modifiers. In general, the solution can be found by iterating this following procedure (Wilcox, 2005):

At step 0, let for all i.

Now set

Repeat until convergence.

4. Performance Measures for Simulation

The design of the simulation permits investigating the statistical properties of the weighted estimator and the two alternative estimators in both an unconditional and a conditional analysis. The conditional analysis considers only the replicates with the influential value, and the unconditional analysis considers all the replicates whether or not an influential value is present. The performance criteria include detection measures as well as the quality measures of bias, variance, and mean square error.

4.1 Detection Measures

The performance assessment includes the following detection measures, which assume there is one known influential value in the population. The measures will be calculated for Clark Winsorization and robust M-estimation:

- Hit rate = the percentage of samples in which the known influential value was detected.
- Type I error rate = the percentage of observations that were not the known influential value but were designated as influential (false positive).
- Type II error rate = the percentage of samples in which the known influential value was not detected (false negative).

The hit rate and type II error rate are both calculated over all samples containing an influential value. The type I error rate is averaged over all 100,000 samples for the unconditional analysis and over just the samples with the influential values for the conditional analysis. Note that the type I error rate will include units which only have a small adjustment made to them, and as such may overstate the degree to which false positives are problematic.

4.2 Relative Bias

For the definitions, we let

Untreated estimate of total *Y* based on replication .

Winsorization estimate of total *Y* based on replication .

Robust M-estimation estimate of total *Y* based on replication .

For the following definitions, we establish the subscript , where setting to represents the untreated estimator, represents the Clark Winsorization estimator, and represents the robust M-estimation estimator. Let denote the set of all indices of all 100,000 replicate samples and the set of all indices of sample replicates that contain the influential value. Further denote the size of a set as .

Define the relative bias of estimator X for the unconditional analysis as

Define the relative bias of estimator X for the conditional analysis as

4.3 Relative Root Mean Square Error

We define the relative root MSE (RRMSE) of estimator X for the unconditional analysis as

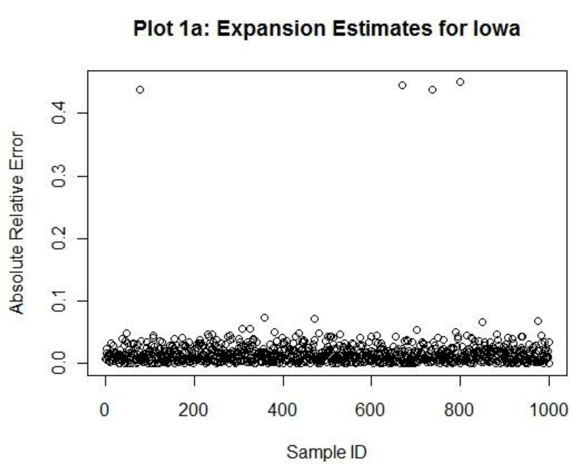
Similarly, we define the RRMSE of estimator X for the conditional analysis as

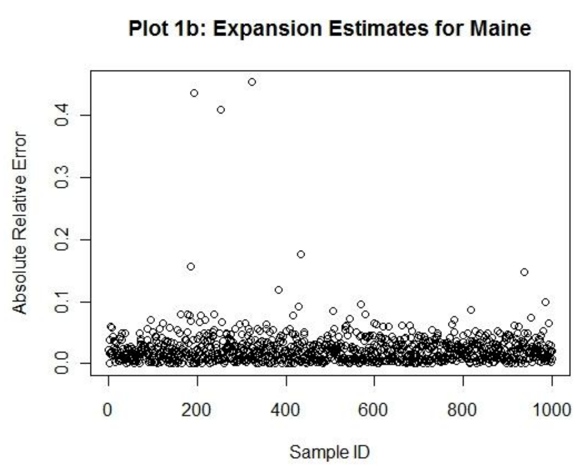
5. Results

We used Iowa and Maine in our simulation because they each had an influential value.

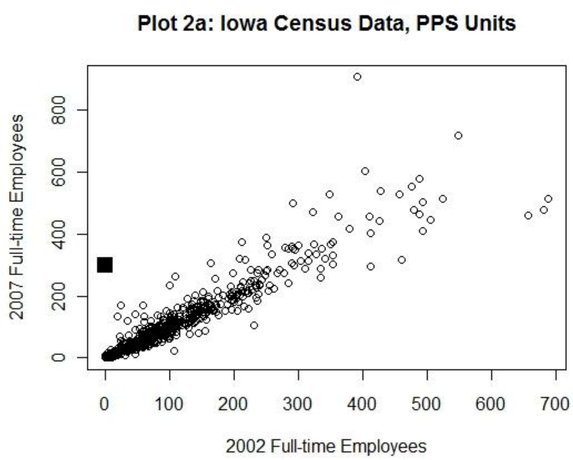
5.1 States with an Influential Value

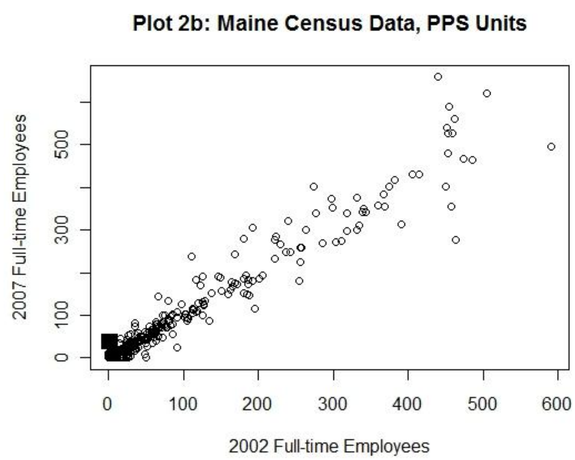
In Plots 1a and 1b, we see the absolute relative error, , of each of our samples for Iowa and Maine. For both Iowa and Maine, we see the majority of samples with less than 10 percent absolute relative error, and a handful of samples with over a 40 percent absolute relative error. Note that Iowa shows two distinct lines; this suggests a single influential value. Maine is similar, although the line near 0 is less well defined which suggests that we may find some moderately unusual values. Investigation of the data showed that our suspicions were correct, with Iowa and Maine each having only one influential value causing the large errors. Note that while not visible from these plots, whenever an influential value was present the error incurred was a considerable overestimation of the true value.



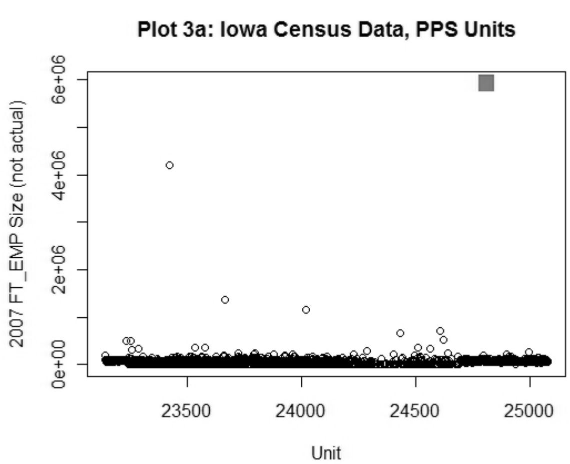


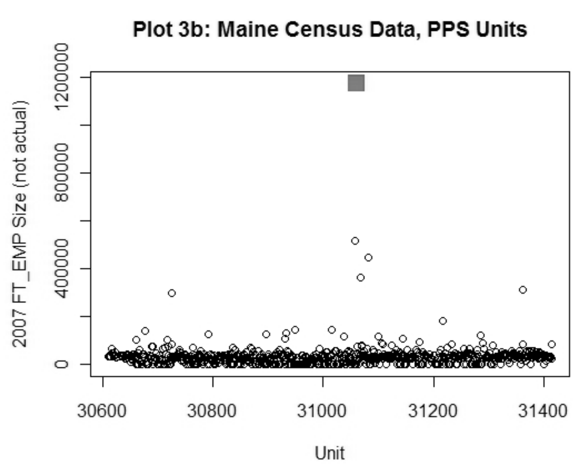
Plots 2a and 2b show the values of full-time employees for all units in both the 2002 and 2007 CoG-E for Iowa and Maine. The data display a strong linear relationship, which validates our use of prior year data as auxiliary variables. In each plot, we denote the influential value identified from the previous two plots as a large square. In the case of Iowa we would likely have looked at this unit even if we started from this plot, but the influential value from Maine does not stand out much from the other units.





Plots 3a and 3b show us the weighted values of all PPS units in the universes for Iowa and Maine. We looked at these plots to see if our assumption that our influential values were unusually large when compared to all other weighted values was valid. Again, the influential values we identified from Plots 2a and 2b are marked with a large square. Note that these units have considerably larger weighted values than the rest of the universe. Iowa is an exception in that there is another unit with a large weighted value which we did not notice in our original analysis. In Maine, we removed a unit from our plot which was due to a large imputed value; we chose to remove it as we are trying to focus on reported data as much as is possible.





5.2 Performance results

We compare our robust estimators against each other and against non-robust estimators in Tables 1 and 2. We only display winsorization results using the robust regression estimate of ; the simple mean and median estimators perform almost but not quite as well. We also looked at calculating our winsorized estimates at two levels: by individual sampling strata and by all units in a state. The results we found did not lead us to believe that there was much to be gained by calculating estimates at the stratum level so we prefer calculating at the state level as it reduces the number of cases of insufficient units for analysis and simplifies the programming. M-Estimation is the clear winner in this analysis, although we suspect that may be due to the very linear cases we chose. In addition, we want to note that in the conditional analysis our RRMSE is due almost entirely to the relative bias of the estimators.

Table 1a: Performance measures for Iowa estimates

Measure	Replicates	True Value	Expansion	GREG	Winsorization: Full State	Winsorization: Individual Strata	M-Estimation
Estimated Value	Full	110,639	110,610	111,797	109,875	109,621	109,223
RRMSE		NA	3.87	4.56	2.50	2.53	2.10
Relative Bias		NA	-0.03	1.05	-0.69	-0.92	-1.30
Estimated Value	Conditional	110,639	161,336	135,828	135,602	135,113	111,890
RRMSE		NA	45.86	23.79	22.65	22.19	2.25
Relative Bias		NA	45.82	22.77	22.56	22.12	1.08

Table 1b: Performance measures for Maine estimates

Measure	Replicates	True Value	Expansion	GREG	Winsorization: Full State	Winsorization: Individual Strata	M-Estimation
Estimated Value	Full	46,347	46,335	46,176	45,984	45,833	46,217
RRMSE		NA	9.94	4.99	5.53	5.56	2.80
Relative Bias		NA	-0.02	-0.37	-0.78	-1.11	-0.56
Estimated Value	Conditional	46,347	65,010	54,365	55,913	55,648	47,882
RRMSE		NA	40.38	18.4	20.85	20.27	8.56
Relative Bias		NA	40.27	17.3	20.64	20.07	3.02

Table 2: Simulation Detection Rates

	Iowa		Maine	
	Winsorization: Full State	**M-Estimation**	**Winsorization: Full State**	**M-Estimation**
Hit Rate	100.0%	100.0%	100.0%	100.0%
Type I Error	2.6%	2.4%	4.4%	0.2%
Type II Error	0.0%	0.0%	0.0%	0.0%

6. Summary and Conclusions

In our analysis, we found major reduction in both relative bias and RRMSE in the presence of an outlier for our outlier robust methods compared against the basic expansion estimator; GREG also showed a sizable improvement over the expansion estimator. The increased bias of these robust methods stays acceptably small when we look at samples where there is no outlier. As such, we can say with confidence that we should be using an outlier robust method for our estimation. In addition, we find that the RRMSE of our robust estimators is due primarily to the relative bias.

We found M-Estimation to be, by far, the superior method for estimation in our study. That said, we looked at two examples where there appeared to be a very strong linear relationship and only one distinct outlier. It is possible that our examples are very well suited to M-Estimation; further analysis of less ideal situations needs to be done before we can fully recommend M-Estimation for production use. In addition, the conceptual simplicity of Winsorization leaves us wanting to further consider its use.

We still have many different cases to analyze before these methods are fully ready for implementation. We need to look at populations with two or more influential values as well as ones that do not show as strong of a linear relationship as those that we looked at here. Since change is an important consideration for users of these estimates, we need to investigate how these methods will impact estimates of year-to-year change. We also would like to look into methods which are better able to treat observations that are “too low” or which only modify an observation if doing so will yield a large reduction in the RRMSE.

Acknowledgements

The authors thank Yang Chang for many helpful conversations and suggestions. The reviews by Lisa Blumerman, Carma Hogue, and Lynn Weidman are appreciated very much.

References

Beaumont, J.-F. And Alavi, A. (2004) “Robust Generalized Regression Estimation.” *Survey Methodology*, 30, 2, 195-208.

Chambers, R. L., Kokic, P., Smith, P. And Cruddas, M. (2000) “Winsorization for Identifying and Treating Outliers in Business Surveys.” *Proceedings of the Second International Conference on Establishment Surveys.* Statistics Canada. Ottawa, Canada. 717-726.

Clark, M. (1995) “Winsorization Methods in Sample Surveys.” Masters Thesis. Department of Statistics. Australia National University.

Mulry, M. H., Oliver, B., and Kaputa, S. (2011) "Detecting and Treating Influential Values in the Monthly Retail Trade Survey via M-estimation & Clarke Winsorization." Presentation at the Statistics Canada/U.S. Census Bureau Interchange, Suitland, MD, March 28-29, 2011. Center for Statistical Research and Methodology. U.S. Census Bureau. Washington, DC.

Mulry, M. H. and Oliver, B. (2009) "A Simulation Study of Treatments of Influential Values in the Monthly Retail Trade Survey." *JSM Proceedings*, Survey Research Methods Section. American Statistical Association. Alexandria, VA. 2979-2993.

Mulry, M. H. and Feldpausch, R. (2007a) "Investigation of Treatment of Influential Values." *Proceedings of the Third International Conference on Establishment Surveys*. American Statistical Association. Alexandria, VA.

Mulry, M. H. and Feldpausch, R. (2007b) "Treating Influential Values in a Monthly Retail Trade Survey." *Proceedings of the Survey Methods Section, SSC Annual Meeting.* Statistical Society of Canada. Ottawa, Ontario, Canada.

Wilcox, R. (2004) *Introduction to Robust Estimation and Hypothesis Testing, 2nd Edition.* Elsevier Academic Press. Burlington, MA.