Setting M-estimation Parameters for Detection and Treatment of Influential Values

Mary H. Mulry, Broderick Oliver, Stephen Kaputa, Katherine J. Thompson[1]
U.S. Census Bureau, Washington, DC 20233

Abstract

Previous research on the use of M-estimation methodology for detecting and treating verified influential values in economic surveys found that initial parameter settings affected its effectiveness. The study relied on simulated data designed to reflect the population properties for two industries in the Monthly Retail Trade Survey (MRTS), but the approach to determining settings for the initial parameters used an empirical analysis that does not generalize well. The need to expand the application of the M-estimation methodology to all the industries in the MRTS and the Monthly Wholesale Trade Survey stimulated the development of a more general methodology that uses historical data to determine the initial parameter settings. This paper discusses the effectiveness of several methods for setting initial parameters, including the important initial value of the tuning constant.

Key words: Outlier, economic surveys

1. Introduction

Recent research demonstrated that M-estimation (Beaumont 2004, Beaumont and Alavi 2007) is effective in identifying influential values, and its algorithm has a flexibility that makes it suitable for application in the Census Bureau economic surveys (Mulry, Oliver, and Kaputa 2012a, 2012b, 2013). An observation is considered influential if its value is correct but its weighted contribution has an excessive effect on the estimated total or period-to-period change. Failure to "treat" such verified influential observations may lead to substantial over- or under-estimation of survey totals, which in turn may lead to overly large increases or exceedingly small decreases in estimates of change. This paper compliments other work on performance measures for M-estimation and extends previous work on the effect of the parameter settings (Mulry, Oliver, and Kaputa 2012a, 2012b, 2013).

In general, business populations are highly skewed. Sample designs are consequently highly stratified, and the sampling rates tend to be higher in the strata with the larger units than in the strata with the smaller units. Typically, economic surveys have a stratified sample design based on major industry with further stratification based on a unit-level size measure such as annual sales, annual payroll, or total assets.

Economic surveys publish totals and period-to-period change estimates. Influential values are examined with respect to their weighted impact on the total. If the total levels vary greatly by period, the change estimates are affected accordingly. When an influential value is detected, the current mitigation strategies depend on whether the subject matter experts believe the observation is a one-time phenomenon or a permanent shift. If the influential value appears to be an atypical occurrence for the business, then the influential observation is replaced with an imputed value. If

[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.

the influential value persists, indicating a permanent change, then methodologists make adjustments to its sampling weight to reflect the change.

Research by Mulry, Oliver, and Kaputa (2012a, 2012b, 2013) found M-estimation appeared suitable for an automated statistical procedure for detecting and treating influential values in a monthly economic survey. The studies found the method retains identified observations in the estimates while assuring their contribution does not have an excessive effect on the monthly totals or an adverse effect on the estimates of month-to-month change. The research used a simulation methodology that generates realistic stable time-series data for the population, enabling the evaluation of the M-estimation method over repeated samples. The findings were that it performed well regarding measures such as relative bias, relative root mean squared error, and number of false detections.

However, the studies highlighted the importance and sensitivity of the M-estimation algorithm to parameter settings, particularly in the event of a single influential value. One aspect of particular interest is the range of values that the methods designate as influential, called the detection region. Values located within this detection region are modified ("treated") to minimize the mean squared error (MSE) of the estimate of the total. M-estimation requires a data model that has few assumptions, but determines the influential value detection region using a highly parameterized algorithm. Consequently, the effectiveness of M-estimation for finding a data-appropriate detection region is highly dependent on the input parameters. The M-estimation procedure can be used to replace a subjective procedure performed by analysts. However, valid initial parameter settings or guidelines for determining such settings must be provided.

This paper proposes methods for setting initial parameters for the M-estimation algorithm and explores their effectiveness via a simulation study with data generated to be a realistic representation of two industries in the MRTS. Also included is a discussion of how the parameter settings relate to the effectiveness of M-estimation in several scenarios for influential values. The paper concludes with an empirical analysis that applies the methods for setting parameters to other industries in the Monthly Retail Trade Survey (MRTS) and expands to consider data from industries in the Monthly Wholesale Trade Survey (MWTS).

2. Method

Before a description of the M-estimation method, which follows Mulry, Oliver, and Kaputa (2012a, 2012b), we first introduce the notation. For the i^{th} business in a survey sample of size n for the month of observation t, Y_{ti} is the collected characteristic (e.g., revenue), w_{ti} is its survey weight (which may or may not be equivalent to the inverse probability of selection), and X_{ti} is a variable highly correlated with Y_{ti}, such as previous month's revenue. The monthly total Y_t is estimated by $\hat{Y}_t$ defined by $\hat{Y}_t = \sum_{i=1}^{n} w_{ti} Y_{ti}$.

For ease of notation, we suppress the index for the month of observation t in the remainder of this section. In many economic surveys such as the MRTS and the MWTS, the survey weight w_{ti} is the design weight with a few individual units' weight adjusted because they are births, deaths, or for item subsampling. No weight adjustments are performed for missing data treatment because imputation is used instead. However, in economic census years, weights may be adjusted to improve coverage.

M-estimators (Huber 1964) are robust estimators that come from a generalization of maximum likelihood estimation. The application of M-estimation examined in this investigation is

regression estimation. The M-estimation technique proposed by Beaumont and Alavi (2004) uses the Schweppe version of the weighted generalized technique (Hampel et al. 1986, p. 315 – 316). The estimator of the total using this approach is consistent for a finite population since it equals the finite population total when a census is conducted (Sarndal et al. 1992, p. 168).

Briefly, the method estimates $\hat{B}^M$, which is implicitly defined by

$$\sum_{i \in S} w_i^*(\hat{B}^M)(y_i - x_i\hat{B}^M)\frac{x_i}{v_i} = 0 \qquad (2.1)$$

where

$v_i = \lambda x_i$

$w_i^*(\hat{B}^M) = w_i \psi\{r_i(\hat{B}^M)\} / r_i(\hat{B}^M)$

$r_i(\hat{B}^M) = h_i e_i(\hat{B}^M) / Q\sqrt{v_i}$

$e_i(\hat{B}^M) = y_i - x_i\hat{B}^M$

The variable x_i may be a vector, but in our application, it is the previous month's value.
The regression model that we employ does not include an intercept because with retail and wholesale trade, the regression of current month's sales on the previous month's sales tends to go through the origin. Section 4 contains a discussion of the settings for the other parameters used in this investigation.

The role of the Huber function ψ is to reduce the influence of units with a large weighted residual $r_i(\hat{B}^M)$. We use the Type II Huber function ψ, which ensures that all adjusted units are at least fully represented in the estimate. The one-sided Type II Huber function is

$$\psi\{r_i(\hat{B}^M)\} = \begin{Bmatrix} r_i(\hat{B}^M), r_i(\hat{B}^M) \le \varphi \\ \frac{1}{w_i} r_i(\hat{B}^M) + \frac{(w_i - 1)}{w_i}\varphi, otherwise \end{Bmatrix} \qquad (2.2)$$

where φ is a positive tuning constant. Detection of observation *i* as an influential value by M-estimation with the Huber II function occurs when $r_i(\hat{B}^M) > \varphi$.

Solving for $\hat{B}^M$ requires the Iteratively Reweighted Least-Squares algorithm in many circumstances. For certain choices of the weights and variables, the solution is the standard least-squares regression estimator.

In implementing M-estimation, the user has a choice of adjusting the weight of the influential value or adjusting its value. The weight adjustment for the Type II Huber function above has the appealing feature of always being greater than one and is given by

$$w_i^*(\hat{B}^M) = \begin{Bmatrix} w_i, r_i(\hat{B}^M) \le \varphi \\ 1 + (w_i - 1)\frac{\varphi}{r_i(\hat{B}^M)}, otherwise \end{Bmatrix} \qquad (2.3)$$

For an adjustment to the influential value, Beaumont and Alavi (2004) use a weighted average of the robust prediction $x_i\hat{B}^M$ and the observed value y_i of the form

$y_i^* = a_i y_i + (1 - a_i) x_i \hat{B}^M$ where $a_i = \frac{w_i^*(\hat{B}^M)}{w_i}$.

Using numerical analysis, Beaumont (2004) finds an optimal value of the tuning constant φ by deriving and then minimizing a design-based estimator of the mean-square error. The minimization does not require a generating data model that holds for all observations (including the influential value).

3. Setting algorithm parameters

The M-estimation algorithm discussed in Section 2 requires settings for Q, h_i, v_i, the function ψ, and an initial value of the tuning constant φ. In this section, we propose methods for setting the parameters for the M-estimation algorithm discussed in Section 2. We motivate the methods by summarizing our results that illustrate the impact of the parameter settings on the effectiveness of the M-estimation algorithm's ability to detect and adjust an influential value. The observation's survey weight also affects whether the algorithm designates it as an influential value. In our simulations to explore the characteristics of the M-estimation algorithm, we used SAS software developed by Jean-Francois Beaumont (2007). We suggest using the default settings for the parameters $Q=1$ and $h_i = (w_i - 1)\sqrt{x_i}$ but explore the potential impact different settings for the other v_i, ψ, and φ. Table 1 summarizes the parameters for the M-estimation algorithm.

Table 1. M-estimation algorithm parameters

Parameter	Parameter Function	Values
Q	Constant	=1 (default)
h_i	Unit weight	$= (w_i - 1)\sqrt{x_i}$ (default)
v_i	Model error underlying regression estimator	= 1 or x_i
ψ	Huber function	Huber I or Huber II
φ	Tuning constant (determines starting point for detection region)	User provides initial value and program calculates optimal value

3.1 Impact of survey weight w_i

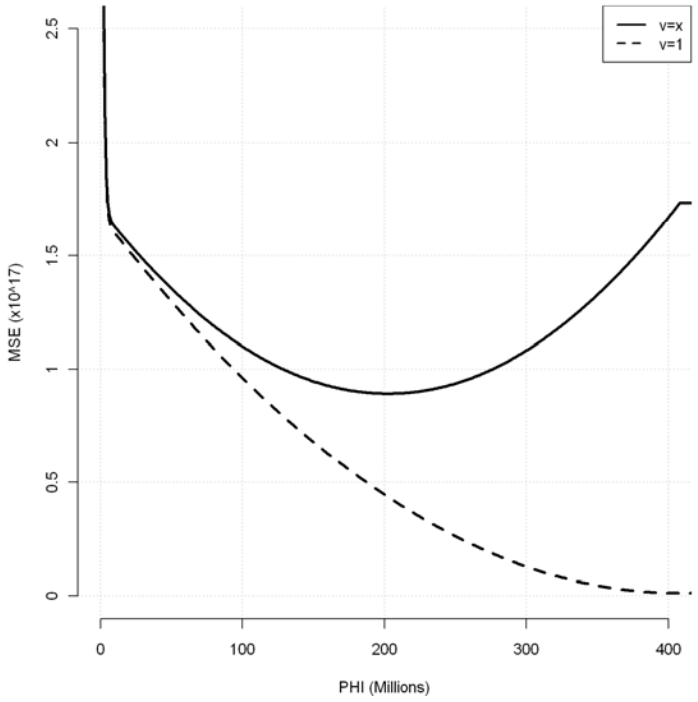
The size of an observation's weight as well as its weighted value both affect whether it will be designated as influential by M-estimation. Typically, the sampling rate for small businesses is lower than for larger businesses because there are more small businesses than larger businesses. Therefore, the smaller businesses typically have higher weights. If two observations have the same unusually high amount of weighted month-to-month change, the M-estimation method is less likely to designate the one with the lower weight as an influential value. Comparably, the weight has a similar effect on the designation of influential if two observations have the same unusually low amount of weighted month-to-month change but different weights. The low observation with the higher weighted is more likely to be designation as influential (Mulry, Oliver, and Kaputa 2012a).

3.2 Parameter v_i

Ideally, the choice of the setting for v_i should be a data-driven decision because v_i essentially specifies the variance of the model errors underlying the regression estimator for M-estimation, denoted by $e_i(\hat{B}^M) = y_i - x_i \hat{B}^M$ in Section 2. The selection of v_i can be the determining factor the effectiveness of the algorithm. We have encountered situations where the M-estimation algorithm produced results for one setting of v_i but not for another, as illustrated in Figure 1. When $v_i = x_i$,

the MSE was a concave function of φ so a minimum MSE is achieved and an adjustment produced for the influential value. However, when $v_i = 1$ is applied to the same sample data, the MSE was a strictly decreasing function of φ so the minimum MSE occurs at the influential value; no adjustment is made because the algorithm fails to detect the influential value. For some insight for the difference in outcome, note that with the defaults $Q = 1$ and $h = (w_i - 1)\sqrt{x_i}$, setting $v_i = 1$ tends to give the residuals for large weighted values of x_i more influence in fitting the M-estimation regression line than when $v_i = x_i$.

Figure 1. M-estimation MSE vs. φ for Industry 1 Month 4 of one sample using Huber II.



To determine of v_i , we suggest fitting regression models for several possible selections and then determining which model gives the best fit for the data through analyses of the residuals. In general, economic surveys have stratified sample designs and units are not selected with equal probability although the selection within strata often is equal probability. Therefore, regression models fit with unweighted economic data usually exhibit heteroscedasticity (unequal error variances).

To illustrate this approach, we focus on MRTS data and the options $v_i = x_i$ and $v_i = 1$. To incorporate the unequal probability selection in parameter computation using sample data, we examine two models the two models below where $\hat{y}_i = w_i y_i$ and $\hat{x}_i = w_i x_i$.

Model 1: $\hat{y}_i = \beta \hat{x}_i + \xi_i,\ \xi_i \sim (0, w_i \sigma^2)$

Model 2: $\hat{y}_i = \beta \hat{x}_i + \xi_i,\ \xi_i \sim (0, w_i x_i \sigma^2)$

Model 1 corresponds to $v_i = 1$ in the M-estimation algorithm, and Model 2 corresponds to $v_i = x_i$.

Model 1 can be fit using SAS Procedure SURVEYREG. Note that Model 2 is equivalent to $\hat{y}_i / \sqrt{\hat{x}_i} = \beta\sqrt{\hat{x}_i} + \xi_i, \xi_i \sim (0, w_i\sigma^2)$, which also can be fit using SAS Procedure SURVEYREG.

We attempted to fit Models 1 and 2 with two consecutive months of data from four industries in the MRTS that did not have influential values. When we examined the residuals, we found that the residual plots for both Models 1and 2 indicated remaining heteroscedasticity, so that neither provided a good model of the data for industries in MRTS.

To learn more, we fit Models 1 and 2 within each stratum. At the stratum level, we found no evidence of heteroscedasticity in the residual plots. Each of the models fit the data well within stratum, as would be expected with the stratified simple random sample design (which assumes equal means and variances within strata). However, the estimate of the coefficient β differed by stratum and the stratum means differed. This analysis validated the sample design fairly well. Unfortunately, a consequence of this sample design validation is that it refutes the assumption of an industry-level data model.

Since neither $v_i = x_i$ nor $v_i = 1$ provided a good model of the MRTS data, we turned our attention to the performance of the algorithm with the two settings. With simulated data modeled from the MRTS (Mulry, Oliver, and Kaputa 2012b), we investigated the number of failures to detect an induced influential value (Type I error) and the number of incorrect detections of observations that were not influential (Type II error). Our investigation found that there is some Type II error when $v_i = 1$ and none when $v_i = x_i$, and the two settings produce about the same results regarding Type I error. We concluded that $v_i = x_i$ was a better choice for our industries.

3.3 Function ψ

The first decision regarding the function ψ whether to use Huber I or Huber II as described in Section 2. After that choice is made, the second decision is whether to use the one-sided version or the two-sided version. The one-sided version by design detects only unusually high influential values while the two-sided version is able to detect both unusually high and unusually low influential values. In previous studies with two MRTS industries, our results were comparable for our Huber I and Huber II when $v_i = x_i$ and when $v_i = 1$.

One thing to consider in choosing the one-sided or two-sided function ψ is that the M-estimation algorithm using a two-sided function ψ may experience some problems with convergence for some scenarios where the second value was too low. The combination of a high influential value and a low influential value causes the algorithm to be less likely to converge. Beaumont (2004) also noted some problems with convergence in his simulations in this situation.

In our previous research (Mulry, Oliver, and Kaputa 2012a), we found that when a sample contains both unusually high and unusually low influential values and the M-estimation algorithm does not converge, no adjustment is probably the best choice. The reason is that the unusual values counterbalance each other in a manner that introduces minimal bias. Therefore, the failure of the algorithm to identify the influential values is not necessarily a handicap. However, if the researcher is not interested in detecting and adjusting unusually low influential values, one-sided version of the function ψ tends to avoid problems with convergence of the algorithm.

3.4 Tuning constant φ

The effectiveness of the M-estimation algorithm is sensitive to the choice of the initial tuning constant φ. The initial φ determines the lower boundary of the detection region, particularly when

there is only one influential value in a sample. In simulations of samples with two high influential values, we found that when we held one influential value fixed and let the second one vary, the detection region for the second one did not appear sensitive to the initial φ.

The goal when setting the initial tuning constant φ is to select a value that is high enough to avoid falsely detecting natural variation as influential, but low enough to detect truly influential values. This is a very delicate balancing act because setting the initial φ too high may result in the algorithm failing to detect influential values that are lower than the initial φ. When none of the values in the sample is larger than the initial φ, the algorithm runs for one iteration and then stops. In this circumstance, the MSE is a constant function in a neighborhood of the initial φ, and the algorithm continues to run only when it detects a change in the MSE in the proximity of the initial φ (Mulry, Oliver, and Kaputa 2013).

On the other hand, setting the initial φ is too low causes the algorithm to give the influential designation to observations not considered influential. This occurs because the algorithm achieves a minimum MSE when there is no influential value by trimming about 0.05 percent of the observations for a very small reduction in the MSE. In an ongoing survey, an initial φ that is too low may also cause convergence problems in a month following an adjustment because the unit returns to its level two months earlier and now appears unusually low. In some cases, both one-sided and two-sided functions ψ have convergence problems (Mulry, Oliver, and Kaputa 2012a).

We explore methods for setting the initial φ in two steps. First, we conduct a simulation study with data generated as a stationary time series so that the evaluation of the performance of the proposed methods is not confounded with seasonal effects. Next, we examine the better methods identified in the first study through an empirical study with data from the MWTS.

4. Simulation study

Our simulations explore three methods for choosing a setting the initial tuning constant φ when an unusually high influential value is present. The proposed options have the potential to avoid detecting natural variation as influential values. In each option, the data used to calculate the initial φ come from the preceding measurement period. Calculating the options requires fitting a regression line where the independent variable is the previous month and the dependent variables is the current month. We explore whether using weighted least squares or weighted robust regression produces better performance. We use the least median of squares (LMS) robust regression method.

The three options are:

- Standard Deviation Method. Set initial φ equal to the product of a factor k and the robust standard error of the residuals, e.g. initial $\varphi = k \times$ (standard error of robust regression residuals). Use normal distribution to determine two options for the value of k, set $k_1 = Z_{1/1,000,000} = 4.753424$ (*low*) and $k_2 = Z_{1/10,000,000,000} = 6.361341$ (*high*). We chose the values of k through trial and error.
- Resistant Fences Method. Set initial φ e by using resistant fences. Use percentiles of residuals to set initial φ where initial φ = quartile(75) +k(quartile(75) –median). To be comparable to the Standard Deviation Method, use $k_j = \frac{z_\alpha - z_{0.25}}{z_{0.25}}$ with α = 1/1,000,000 resulting in $k_1 = 6.047437$ (*low*), and α = 1/10,000,000,000 resulting in $k_2 = 8.431338$ (*high*). When setting k for resistant fences, the relevant measure is the *some-outside rate per sample*, the probability that a sample has an observation is flagged as outlying by chance. Hoaglin and Iglewicz (1987) found that the some-outside rate per sample for

normally-distributed data with a sample size of $n \geq 200$ was approximately 10% for k = 2.2 and 5% for k = 2.4. Extrapolating that an increase in k of 0.2 reduces the some-outside rate per sample by half, an estimate of the some-outside rate is 3/1,000,000 for k_1 = 6.361341 (*low*) and 8/10,000,000,000 for k_2 = 8.431338 (*high*).

- Bootstrap Method. Set initial φ using a stratified bootstrap. Repeatedly draw bootstrap samples (Efron 1981). For each sample, choose the observation corresponding the 99.00^{th} percentile for the *low* value and the observation corresponding to the 99.99^{th} percent for the *high* value. Then average the low values over all replicates and the high values over all replicates. Often both percentiles were equal to the maximum value in the sample.

If the residuals have a normal distribution, then the Standard Deviation and Resistant Fences methods produce approximately the same values of the initial φ. However, if the distribution of the residuals is not normal, then the values of the initial φ produced by the two methods are different. The value of the initial φ produced by the Bootstrap method will be different from the values produced by the Standard Deviation and Resistant Fences methods regardless of whether the residuals have a normal distribution.

4.1 Simulated data

To assess the performance of the three options, we conduct a simulation study similar to the method used by Mulry, Oliver, and Kaputa (2012a). The simulated population data presents "realistic" monthly sales estimates, modeled from two industries with different natures. One that we refer to as SIM-R1 has monthly sales of approximately 46.1 billion and one of the most volatile patterns for influential values. The other that we refer to as SIM-R2 has a more stable pattern and has monthly sales of approximately 2.5 billion. The sample sizes in our simulations are 1,161 for SIM-R1 and 147 for SIM-R2.

The models used to simulate populations for SIM-R1 and SIM-R2 use the MRTS data for these industries. Recall that the MRTS is a stratified sample, with strata defined by unit size within industry where the measure of size is sales. To obtain realistic level estimates, we apply the nonparametric resampling algorithm described in Thompson (2000) by industry-strata to empirical MRTS data to obtain Month 1 data, thus ensuring that the strata means are different and the industry totals equal the survey estimates. Then, we generate six additional months of the population data for each sampling stratum h in the industry using ARMA modeling to form a stationary series for that stratum, so that $\hat{y}_{hi,t} = \beta_h \hat{y}_{hi,t-1} + \varepsilon_{hi}, \varepsilon_{hi} \sim (0, w_{ti}\sigma^2_{hit}), t > 1$. Therefore, each of the two populations is a stationary series within strata, but not at the industry level. Since the time series is stationary, the strata-level means are approximately the same over time although in practice the strata-level means may vary over a similar period. Using a stationary series avoids the possibility of a trend confounding the effects of the influential values.

Once we have constructed the time series for the population for an industry, we select 1000 samples and induce an unusually high influential value in Month 4 (Mulry, Oliver, and Kaputa 2012a). For this conditional analysis, we use data from Month 2 to calculate each of the three options for the initial φ and use each on in an application of the algorithm in Month 3. Then we re-calculate the initial φ for each of the three options using data from Months 3 for application in Month 4. We repeat the calculation the initial φ for each month using data from the preceding month.

4.2 Performance results

In this section, we examine the simulation results regarding the performance of M-estimation with different options for the parameter settings the quality of the estimates they produce. In Table 2, we focus on performance measures for estimates of totals and month-to-month change.

Table 2. Performance of estimates of total for Industry SIM-R2 with high and low options of three methods of setting the initial φ when an unusually high influential value is induced in Month 4 using weighted robust regression with 1,000 replications and the Huber II function.

method	level	month	Average number of false positives	RB of untreated total	RB of treated total	RRMS of untreated	RRMS of treated
St. dev.	Low	3	0	0.1421	0.1421	2.5660	2.5660
St. dev.	High	3	0	0.1421	0.1421	2.5660	2.5660
R-Fence	Low	3	0.089	0.1421	0.1404	2.5660	2.5661
R-Fence	High	3	0.006	0.1421	0.1419	2.5660	2.5659
B-strap	Low	3	1.872	0.1421	0.1112	2.5660	2.5664
B-strap	High	3	1.015	0.1421	0.1250	2.5660	2.5682
St. dev.	Low	4	0	18.6411	9.2570	18.8149	9.6018
St. dev.	High	4	0	18.6411	9.2570	18.8149	9.6018
R-Fence	Low	4	0	18.6411	9.2568	18.8149	9.6016
R-Fence	High	4	0	18.6411	9.2570	18.8149	9.6018
B-strap	Low	4	0	18.6411	9.2533	18.8149	9.5984
B-strap	High	4	0	18.6411	9.2550	18.8149	9.6000
St. dev.	Low	5	0	0.0950	0.0950	2.5565	2.5565
St. dev.	High	5	0	0.0950	0.0950	2.5565	2.5565
R-Fence	Low	5	2.219	0.0950	-0.0101	2.5565	2.6365
R-Fence	High	5	0.249	0.0950	0.0839	2.5565	2.5526
B-strap	Low	5	0	0.0950	0.0950	2.5565	2.5565
B-strap	High	5	0	0.0950	0.0950	2.5565	2.5565
St. dev.	Low	6	0	0.1698	0.1698	2.5760	2.5760
St. dev.	High	6	0	0.1698	0.1698	2.5760	2.5760
R-Fence	Low	6	0.025	0.1698	0.1692	2.5760	2.5761
R-Fence	High	6	0.004	0.1698	0.1697	2.5760	2.5760
B-strap	Low	6	1.473	0.1698	0.1423	2.5760	2.5726
B-strap	High	6	0.321	0.1698	0.1620	2.5760	2.5752
St. dev.	Low	7	0	0.1359	0.1359	2.5856	2.5856
St. dev.	High	7	0	0.1359	0.1359	2.5856	2.5856
R-Fence	Low	7	0.096	0.1359	0.1340	2.5856	2.5859
R-Fence	High	7	0.002	0.1359	0.1359	2.5856	2.5856
B-strap	Low	7	1.83	0.1359	0.1095	2.5856	2.5838
B-strap	High	7	1.112	0.1359	0.1183	2.5856	2.5849

For the M-estimation algorithm, we used SAS software developed by Jean-Francois Beaumont (2007). Calculating the options for the initial φ requires fitting a regression line. We considered two regression methods, weighted least squares and weighted robust regression.

Table 2 shows the performance results when an unusually high influential value is induced in Month 4 with data from Industry SIM-R2. We induced the influential value by selecting a unit from a small business stratum with a weight of 50 and adding 8 million to its unweighted value. The performance measures we examine are the average number of observations detected as influential that were not induced influential values (*false positives*), the relative bias(*RB*) and the relative root mean square error (*RRMSE*) of the estimate of total with the values designated as influential adjusted (*treated*) and with no adjustments (*untreated*).

Both the high and low levels of the Bootstrap method did not perform as well as the Standard Deviation and Resistant Fences methods under the criteria of lowest false positive detections of influential values and lowest RB and RRMSE. Within the Standard Deviation and Resistant Fences options, the high level of both produced the best performances on all three criteria as shown for Industry SIM-R2 in Table 2. The performance was better when using residuals from weighted robust regression than for weighted least squares regression although the results for the latter do not appear in this paper. Performance results data for Industry SIM-R1 were similar to those for Industry SIM-R2 even though Industry SIM-R1 is smaller with residuals that have a more normal distribution while Industry SIM-R1 is larger with residuals that have a heavy-tail distribution. Therefore, we selected the high level of the Standard Deviation and Resistant Fences options using robust regression for further study in the empirical analysis.

5. Empirical analysis for initial φ

Now we apply what we have learned in Section 4 about the performance of the different approaches for setting the parameters to historic data for four MWTS industries. The historic MWTS data is subject to seasonal effects whereas the simulated data was a stationary series. We use the results of this analysis to calculate the parameters so that we can run M-estimation in a side-by-side experiment with the current method of detecting and treating influential values.

We continue to explore more than one option for setting the initial φ to be sure the best methods in the simulation data appear to be the best methods for a wider range of industries. We examine the high options of both the Standard Deviation and the Resistant Fences methods, but vary the data used in the calculations with the following four alternatives summary statistics:

1) Calculate the initial φ each month with data from the previous month, as in the simulations in Section 4.
2) Calculate the initial φ at the beginning of each year using data from each of the 12 months of the previous calendar year and taking the maximum of the 12 values of φ.
3) Calculate the initial φ at the beginning of each year using data from each of the 12 months of the previous calendar year and taking the mean of the 12 values of φ.
4) Calculate the initial φ each month but uses the data from the same month in the previous year.

The empirical analysis uses 25 consecutive months of observed data from each of two MWTS industries, which we call OBS-W1 and OBS-W2. Industries OBS-W1 and OBS-W2 are not related to the industries that provided the basis for generated the simulated data, SIM-R1 and SIM-R2, used in the previous analysis. Industry OBS-W1 tends to be more volatile than Industry OBS-W2. The first 13 months of data provided the data to calculate the initial φ using the different methods for application in the last 12 months. The sample size for Industry OBS-W1

was 164, and the average of its estimated total sales during the last 12 months was about 2.5 billion. The sample size for Industry OBS-W2 was 86, and the average of its estimated total sales was about 38 billion. Because of space limitations, we show results for only Industry OBS-W1, but the results for Industry OBS-W2 are similar.

The performance measures for evaluating the combinations of the two methods and four alternatives are the number of observations designated as influential and the reduction in estimated MSE obtained by the adjustments. Remember that since we are using real data instead of simulated data, we do not have a true population total to use in the calculation of the true MSE and must settle for an estimated MSE.

Figure 2 shows the initial φ for each of the four options for Months 14 to 25 when using the Standard Deviation and Resistant Fences methods with data from Industry OBS-W1. The values of the initial φ from the Standard Deviation method tended to be higher than the initial φ from the Resistant Fences method across all 12 months and all four alternatives.

Figure 2. Value of initial φ (in millions) for MWTS Industry OBS-W1 for four options for the Standard Deviation and Resistant Fences methods by month

The effects of lower values of the initial φ from the Resistant Fences method are present in Tables 3 and 4. Table 3 shows the performance results of the application of the four alternatives using the Standard Deviation high option to data from Industry OBS-W1. Table 4 is analogous to Table 3 except the Resistant Fences high option method is used. Within the Standard Deviation and the Resistant Fences methods, the alternative that uses data from the *same month in the previous year* to calculate the initial φ results in the lowest number of observations designated as influential. When both methods make adjustment in Month 19, the Standard Deviation method adjusts one unit while the Resistant Fences method adjusts four units. Interestingly, the reduction in the estimated MSE is somewhat higher for the Standard Deviation method at 14.7% than the Resistant Fences method at 13.6%.

Figures 3 and 4 aid in assessing the performance of the alternative that uses data from the same month in the previous year. Both figures show the weighted sample observations not included with certainty for all 25 months in gray [Note: certainty observations are not considered in outlier detection and treatment in M-estimation. See equations 2.2 and 2.3]. Keep in mind that the scale of the graphs is determined by the *weighted observations* (receipts value) while consideration for an influential value designation depends on a comparison between the *weighted regression residuals* and the initial φ. The algorithm focuses on weighted month-to-month change for a unit rather than the level of the weighted observation.

Table 3. Empirical analysis performance measures for four options of setting the initial φ using the Standard Deviation high option in Industry OBS-W1

	Method for initial φ							
	Previous month		Previous year max		Previous year mean		Previous year month	
Month	Number adjusted	Reduction in MSE (%)	Number adjusted	Reduction in MSE (%)	Number adjusted	Reduction in MSE (%)	Number adjusted	Reduction in MSE (%)
14	0	0.0	0	0.0	0	0.0	0	0.0
15	0	0.0	0	0.0	0	0.0	0	0.0
16	0	0.0	0	0.0	0	0.0	0	0.0
17	0	0.0	0	0.0	0	0.0	0	0.0
18	0	0.0	0	0.0	0	0.0	0	0.0
19	1	14.7	1	14.7	1	14.7	1	14.7
20	0	0.0	0	0.0	0	0.0	0	0.0
21	0	0.0	0	0.0	0	0.0	0	0.0
22	0	0.0	0	0.0	0	0.0	0	0.0
23	0	0.0	0	0.0	0	0.0	0	0.0
24	0	0.0	0	0.0	0	0.0	0	0.0
25	5	43.7	0	0.0	0	0.0	0	0.0

Table 4. Empirical analysis performance measures for four options of setting the initial φ using the Resistant Fences high option in Industry OBS-W1

	Method for initial φ							
	Previous month		Previous year max		Previous year mean		Previous year month	
Month	Number adjusted	Reduction in MSE (%)	Number adjusted	Reduction in MSE (%)	Number adjusted	Reduction in MSE (%)	Number adjusted	Reduction in MSE (%)
14	4	37.0	4	37.0	4	37.0	4	37.0
15	3	1.5	0	0.0	3	1.5	0	0.0
16	3	0.1	0	0.0	3	0.1	0	0.0
17	0	0.0	0	0.0	0	0.0	0	0.0
18	0	0.0	0	0.0	0	0.0	0	0.0
19	4	13.6	4	13.6	4	13.6	4	13.6
20	0	0.0	0	0.0	0	0.0	0	0.0
21	0	0.0	0	0.0	0	0.0	0	0.0
22	0	0.0	0	0.0	0	0.0	0	0.0
23	0	0.0	0	0.0	0	0.0	0	0.0
24	0	0.0	0	0.0	2	2.1	2	2.1
25	5	43.7	5	43.7	5	44.2	5	44.2

Figure 3. Standard Deviation method using the same month previous year option with data from Industry OBS-W1

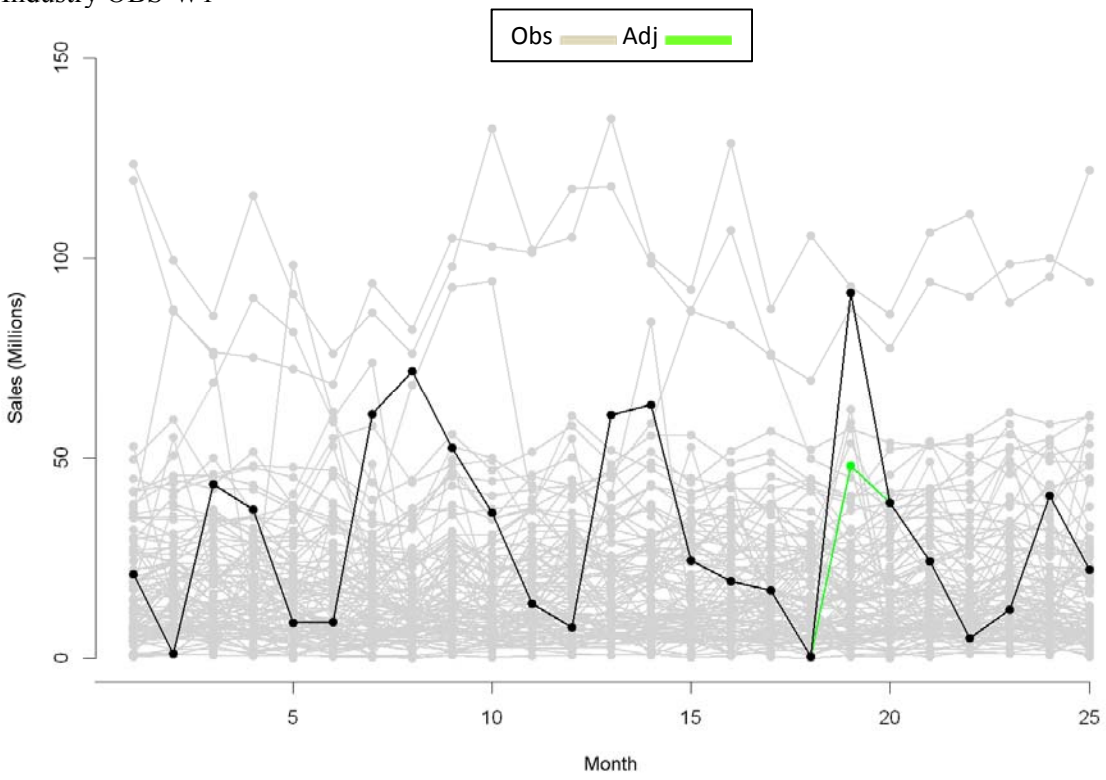
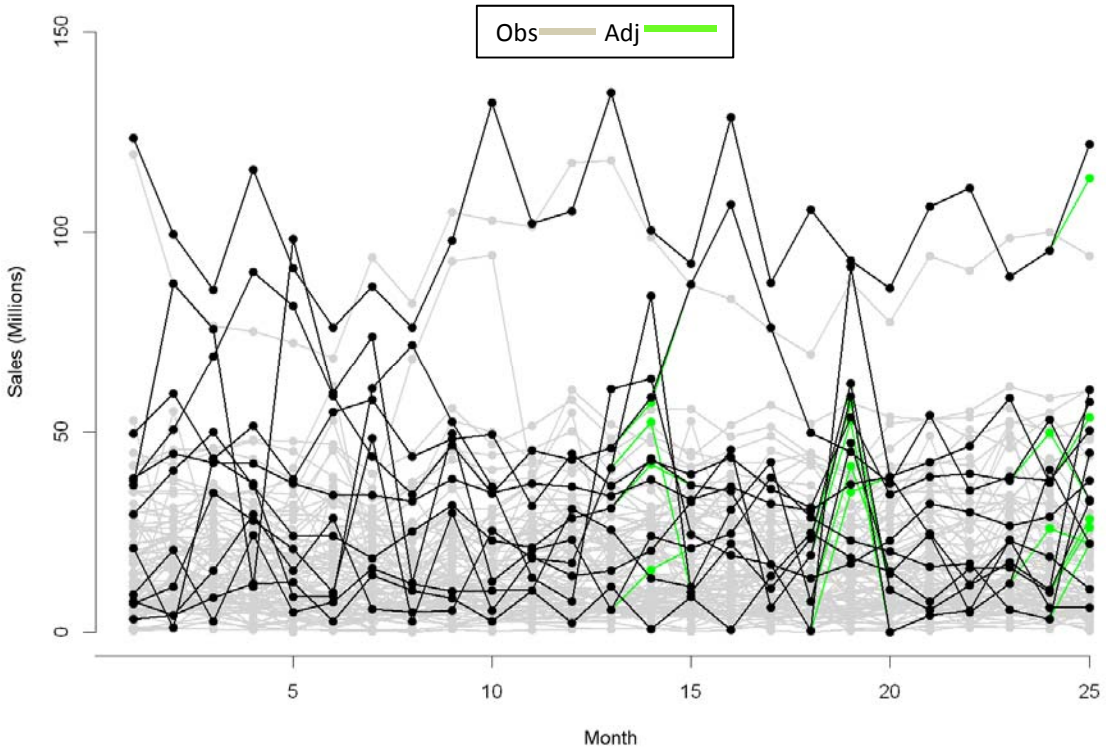


Figure 4. Resistant Fences method using the same month previous year option with data from Industry OBS-W1



In the figures, if the weighted observation for a sample unit is adjusted in one month, its values for all months are shown in black and connected by black lines. The adjusted values are in green with green lines connecting them to their values in the previous and subsequent months.

In Figure 3, we see that the M-estimation algorithm with the initial φ set using the Standard Deviation method designates one weighted observation in Month 19 as influential. The observations for the designated unit across the 25 months are very variable. However, its weighted observation in Month 19 is unusually high when compared to all the other months and particularly when compared to the previous month. In addition, it is one of the highest weighted observations over the 25 months. This adjustment has the effect of reducing the bias in the estimate of total sales to where the estimated bias squared is 8.5% of the estimated MSE.

Now we turn our attention to Figure 4 and see that when the algorithm uses the Resistant Fences method, weighted observations in Months 14, 19, 24, and 25. Many of the adjustments do not appear to make large changes. This is an indication that the algorithm is achieving a minimum estimated MSE by trimming a few observations to achieve a reduction in the variance of the estimated total sales while the estimated squared bias is 24.1% of the estimated MSE.

6. Summary

Using M-estimation to identify and treat influential values in a survey setting is appealing from both methodological and statistical perspectives. The flexibility of weighted M-estimation makes it useful for a wide variety of data models, and our empirical results appear to support the algorithm's robustness to model misspecification. On the other hand, this same flexibility has the disadvantage of introducing some complexity in implementation. First, there are situations when the algorithm has convergence issues, but careful setting of the parameters for the algorithm appears to reduce this problem and sometimes avoids it all together. These convergence issues tend to be more difficult to avoid when the algorithm uses a two-sided function ψ implementation than with a one-sided function. If the lack of convergence is caused by the occurrence of both an unusually high and an unusually low influential value in the same month, then an estimate with no adjustments is justified because the two influential values offset to result in the bias being approximately zero.

In this paper, we explore the basic question of how to develop initial settings for the M-estimation parameters, focusing primarily on economic data applications. The populations that we studied are highly skewed and are consequently highly stratified. Because of this, the assumed data model that we use in our M-estimation application -- a weighted regression model that uses survey weights and the predictor variable as regression weights – is misspecified when applied to population data. Even so, we found several advantages of using this data model over the simpler ordinary least squares (equal variances) model.

Developing an "automatic" method for setting the initial value of the tuning constant φ posed a more challenging problem, especially given the seasonality in our monthly estimates. Since this parameter has the most impact on the performance of the detection of influential values, it is important to provide simple-to-use and data-based methods that are robust. Of all the methods that we considered, the Standard Deviation high option method applied using data for the same month in the previous year yielded the best performance. This combination creates adjustments that reduced bias and achieved the lowest estimated MSE.

The next step is to apply the method in a side-by-side test. We will provide guidelines to the subject matter experts who have the responsibility of reviewing an adjustment proposed by the M-estimation algorithm and deciding on whether to incorporate it in the estimation each month. The dialog with subject matter experts during the test and the application of the algorithm in more industries may lead to refinements, but the basic approach appears very effective.

Acknowledgements

The authors are very grateful to Jean-Francois Beaumont for providing the SAS code for the M-estimation algorithm and for all his advice and consultations. They appreciate the reviews by Eric Slud, Darcy Steeg Morris, Scott Scheleur, William Abriatis, and William J. Davie Jr. very much.

References

Beaumont, J.-F. (2007). personal communication.

Beaumont, J.-F. (2004). "Robust Estimation of a Finite Population Total in the Presence of Influential Units." Report for the Office for National Statistics, dated July 23, 2004. Office for National Statistics, Newport, U.K.

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

Efron, B. (1981). "Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods." *Biometrika.* 68. 589-599.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Werner, S. A. (1986). *Robust Statistics. An Approach Based on Influence Functions.* John Wiley & Sons. New York, NY.

Hoaglin, D. and Iglewicz, B. (1987) "Fine-tuning Some Resistant Rules for Outlier Labeling." *Journal of the American Statistical Association*. American Statistical Association. Alexandria, VA. 83. 1147-1149.

Huber, P. J. (1964). "Robust Estimation of a location parameter." *Annals of Mathematical Statistics*. Institute of Mathematical Statistics. 35. 73-101.

Mulry, M. H., Oliver, B., and Kaputa, S. (2013). "A Note on Setting the M-estimation Tuning Constant in a High Influential Value Scenario." Unpublished manuscript. Center for Statistical Research and Methodology. U.S. Census Bureau. Washington, DC.

Mulry, M. H., Oliver, B., and Kaputa, S. (2012a). "Several Scenarios for Influential Observations and Methods for Their Treatment." *2012 JSM Proceedings*. American Statistical Association. Alexandria, VA. 4015-4029. http://www.amstat.org/sections/SRMS/Proceedings/y2012/Files/304652_73493.pdf

Mulry, M. H., Oliver, B., and Kaputa, S. (2012b). "Study of Treatment of Influential Values in a Monthly Retail Trade Survey" *Proceedings of the Fourth International Conference on Establishment Surveys*. American Statistical Association. Alexandria, VA. http://www.amstat.org/meetings/ices/2012/papers/301892.pdf

Sarndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag. New York, NY.