Outlier Detection and Treatment in the Current Employment Statistics Survey

Julie Gershunskaya and Larry Huff, BLS
J. Gershunskaya, Statistical Methods Division, Office of Employment and Unemployment Statistics, PSB 4985, U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Washington, DC 20212

Key words: robust estimation, one-step M-estimator, ratio estimator, influence curve

Abstract: The Current Employment Statistics (CES) Survey uses a weighted link relative estimator to make estimates of employment at various levels of industry and area detail. The estimates are produced monthly approximately three weeks after the reference date of the survey. Sometimes outliers combined with relatively large probability weights result in influential reporters that cause estimates of smaller domains to be very unstable. An employment figure reported to the survey may be considered typical for a relatively large estimation domain, however, it may be unusual and highly influential for a more detailed industry and area domain. The focus of the current simulation study is to explore the feasibility of using a robust estimation technique in a simple and automated way to detect and treat outliers during the short timeframe allotted for monthly survey processing. Results are evaluated based on the deviation of the estimates from the true population levels.[1]

1. Introduction.

The Current Employment Statistics (CES) is a large-scale Federal-State establishment survey conducted by the Bureau of Labor Statistics. It produces estimates of employment, paid hours and earnings. The estimates are produced monthly, approximately three weeks after the reference period.

National estimates for various industrial levels are produced in the national office while State estimates are produced in State offices. In addition to State total and statewide industry level estimates, States publish estimates for intersections of industrial supersectors and geographical areas such as metropolitan statistical areas (MSA).

[1] Opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

Often a single or a few outlying reports induce much instability in the estimates of smaller domains. The "offenders" do not necessarily report incorrect data: the erroneous reports are being identified during the initial screening conducted by analysts at the national office (see Esposito, Fox, Lin and Tidemann, 1994, McConnell, 1999, for the detail description of screening procedures used in the CES). In many cases, however, a problem arises when an unusual change in reported employment is combined with a relatively high selection weight. The report may look "innocent" when used in the estimates for higher levels of aggregation, however, it may gain an undue influence at lower levels.

State analysts need an objective, relatively simple and automated (that is especially important for the strict time constraints of the production process) procedure to identify and restrain the effect of such reports.

In this paper, we focus on the estimates of employment. In the two following paragraphs, we describe briefly the sample selection process and the form of the estimates used in the CES. In Section 2, we consider the sample influence function for the estimator of the relative employment growth. Models and the forms of the robust estimator of the relative monthly change are considered separately for small and for large samples in Section 3. The form of the estimator for levels is given in Section 4. Empirical evaluation results are presented in Section 5.

The sample for the CES is selected once a year from each State's unemployment insurance (UI) file. The UI accounts (that may include a single establishment or a group of establishments) on the frame are divided into strata by State, major industrial division (also called industrial supersectors) based on NAICS, and employment size class. Nearly optimum sampling allocation targeting the variance of each state-level estimate of employment change at a given cost per State determines the sample size for each stratum.

The CES survey uses a weighted link relative (WLR) estimator at various levels of aggregation: estimate for current month employment level derived by application of the estimate of one month employment growth rate to the previous month estimate: $\hat{X}_t = \hat{X}_{t-1}\hat{\beta}_{t-1,t}$. Once a year, an estimate is benchmarked to a census level figure (that becomes available on a lagged basis): $\hat{X}_{t=1} = X_0^{BMK}\hat{\beta}_{t=1}$. The relative monthly growth rate is estimated from the matched sample $S_{t-1,t}$ (units reporting nonzero employment in both previous and current months) as a ratio of weighted reported employment, current-to-previous months:

$$\hat{\beta}_{t-1,t} = \frac{\sum_{i \in S_{t-1,t}} w_i y_{i,t}}{\sum_{i \in S_{t-1,t}} w_i y_{i,t-1}},$$

where w_i is a selection weight, y_{t-1}, y_t are the values of reported employment at corresponding months. There is no explicit nonresponse adjustment incorporated in the estimator.

2. Influence function for the estimator of one month relative change.

Let (Y_{t-1}, Y_t) be a pair of random variables from distribution $F_{t-1,t}$, and $\theta_t = \frac{\alpha_t}{\alpha_{t-1}}$, the ratio of marginal expectations, is the parameter of interest. For a given population of N units, the quantity

$$\theta_{t,N} = \frac{\sum_{i=1}^{N} y_{i,t}}{\sum_{i=1}^{N} y_{i,t-1}},$$

a ratio of current-to-previous months population employment, is an estimator of θ_t.

Denote $u_i(\theta_t) = \frac{y_{i,t} - \theta_t y_{i,t-1}}{\alpha_{t-1}}$, which is an influence function that measures the sensitivity of the estimator $\theta_{t,N}$ of the parameter θ_t to realized population values $(y_{i,t-1} y_{i,t}), i = 1 \ldots N$.

The estimator $\theta_{t,N}$ can also be viewed as a solution in θ_t of

$$\sum_{i=1}^{N} u_i(\theta_t) = 0 \qquad (2.1)$$

A sample estimate of a population mean $\bar{u}_{t,N} = \frac{1}{N}\sum_{i=1}^{N} u_i(\theta_t)$ is given by $\hat{\bar{u}}_{t,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{w_i}{\bar{w}_n} u_i(\theta_t)$, where n is a sample size and w_i is a probability weight $w_i = 1/\pi_i$, π_i is a probability of selection; $\bar{w}_n = \frac{1}{n}\sum_{i=1}^{n} w_i$.

Let us regard $y_{i,t}^w = \frac{w_i}{\bar{w}_n} y_{i,t}$ and $y_{i,t-1}^w = \frac{w_i}{\bar{w}_n} y_{i,t-1}$, $i = 1 \ldots n$ an observed values of i.i.d. pairs of random variables (Y_{t-1}^w, Y_t^w) under a sample distribution (see, for example, Pfeffermann and Sverchkov, 2003). The influence of a sample unit i on the estimator $\theta_{t,n} = \frac{\sum_{i=1}^{n} w_i y_{i,t}}{\sum_{i=1}^{n} w_i y_{i,t-1}}$ of θ_t can be expressed as

$$u_i^w(\theta_t) = \frac{y_i^w - \theta_t y_{i,t-1}^w}{\alpha_{t-1}} \qquad (2.2)$$

3. Robustified sample estimate of one month relative change.

In order to proceed with the robustification of the estimator $\theta_{t,n}$, we need to introduce a model for the random variables $U_{t,h} = Y_{t,h} - \theta_t Y_{t-1,h}$ in stratum h $(h = 1 \ldots H)$. Let us assume that

$$\begin{aligned} E_P(U_{t,h}) &= \mu_{t,h} \\ Var_P(U_{t,h}) &= \sigma_{t,h}^2, \end{aligned} \qquad (3.1)$$

where expectation and variance are with respect to the population distribution.

First, we consider a model for **small samples**.

Let us assume that

$$\mu_{t,1} = ... = \mu_{t,H} = \mu = 0 \qquad (3.2)$$

Further, noting the way we selected the sample, the probability weights w_i to be inversely proportional to the standard deviations of one-month change across strata, assume

$$w_{i\in h} \propto \frac{1}{\sigma_{t,h}} \qquad (3.3)$$

For the probability weighted $U_t^w = U_t / \pi_i$, therefore, model with respect to the sample distribution is

$$\begin{aligned} E_S\left(U_{t,h}^w\right) &= 0 \\ Var_S\left(U_{t,h}^w\right) &= \sigma_t^2 \end{aligned} \qquad (3.4)$$

Our assumption is that the variance in (3.4) is a constant across strata. Assume also normal density in the middle of the distribution and symmetrical tails. The estimation equation we consider for the robustified estimator is

$$\sum_{i=1}^{n} \psi\left(\frac{w_i u_i(\theta_t)}{\sigma_t}\right) = 0, \quad (3.5)$$

with $\psi(x)$ a bounded increasing odd function, $\psi(x) = x$ for small values of x, e.g. a Huber ψ-function $\psi(x) = \min(\max(x, -c), c)$.

If the model holds, the estimator defined by (3.5) has a smaller mean squared error than the original estimator, while model failure may lead to a bias and increased mean squared error.

We now revise the model for the case of **large sample** relaxing assumptions (3.2), (3.3), and allowing for asymmetrical tails.

Let us consider separately two finite populations: $P_t^{Left} = \left\{i \in P_t \mid u_{i,t} \le 0\right\}$ and $P_t^{Right} = \left\{i \in P_t \mid u_{i,t} > 0\right\}$. Let us assume two separate models analogous to (3.1). For P_t^{Left},

$$\begin{aligned} E_P\left(\left|U_{t.h}\right|\right) &= \mu_{t,h}^{Left} \\ Var_P\left(\left|U_{t,h}\right|\right) &= \sigma_{t,h,Left}^2 \end{aligned}.$$

Similarly, for P_t^{Right},

$$\begin{aligned} E_P\left(U_{t,h}\right) &= \mu_{t,h}^{Right} \\ Var_P\left(U_{t,h}\right) &= \sigma_{t,h,Right}^2 \end{aligned}.$$

Let $\hat{u}_{t,N^L}^{Left} = \sum_{i=1}^{N^L} \left|u_{i,t}\right|$ and $\hat{u}_{t,N^R}^{Right} = \sum_{i=1}^{N^R} u_{i,t}$ be parameters of interest.

We turn to the Kokic and Bell (1994) method and apply it to find Winsorizing cutoffs L_t^{Left} and L_t^{Right} for the estimates of $\hat{u}_{t,N^L}^{Left}$ and $\hat{u}_{t,N^R}^{Right}$.

Let the estimating equation be $\sum_{i=1}^{n} \psi\left(w_h u_{i\in h,t}^*\right) = 0$,

where $u_{i\in h,t}^* = \begin{cases} u_{i\in h,t} - \mu_h^R, & u_{i\in h,t} > 0 \\ u_{i\in h,t} + \mu_h^L, & u_{i\in h,t} \le 0 \end{cases}$, and

and $\psi(x) = \begin{cases} -L^{Left}, & x < -L^{Left} \\ x, & -L^{Left} \le x \le L^{Right} \\ L^{Right}, & x > L^{Right} \end{cases}$

4. Form of the estimates for levels.

We used the following formula for the estimates of the employment level:

$$\hat{X}_t = \sum_{i=1}^{n} y_{i,t} + \hat{\theta}_{t,n}^R \left(\hat{X}_{t-1} - \sum_{i=1}^{n} y_{i,t-1}\right),$$

where $\hat{\theta}_{t,n}^R$ is a robustified estimate of relative employment change.

5. Empirical evaluation.

The study covers the period that starts from the March 2002 benchmark level, the estimates are produced from April 2002 through September 2003. The results at the end of the estimation period (i.e., the estimates for September 2003) are compared to the true population levels (that become available approximately nine months after the reference date). Small sample results (for cells having, on average during the estimation period, 10 to 50 reporting units) are presented in application to the domains defined by the intersection of metropolitan statistical areas (MSA) and major industrial divisions (called industrial supersectors, or SS). Large samples (more than 50 reports, on average during the estimation period) studied at a statewide industrial supersector level.

For small samples, we considered Huber one-step and fully iterated estimators with original sample estimate as initial value for $\hat{\theta}_t$, tuning constant *C=2*, scale was estimated by MAD: $\hat{\sigma}_t = med\left(\left|w_i u_i - med(w_i u_i)\right|\right)/0.67$.

One-step Huber estimator decreases overall root MSE to 74.1% of the original estimator's root MSE; the fully iterated version gives 70.4% of the original root MSE. Both robust estimators work uniformly well across industries (Fig.1).

For a State-by-SS domains, estimates based on the Winsorization algorithm described in section 3 (let us denote it Method 2) are being compared to the Huber-type estimator (Method 1) with *C=6* and scale estimated by

$$\hat{\sigma}_t = Q^{.75}\left(\left|w_i u_i - med(w_i u_i)\right|\right).$$

Overall root MSE decreases to 88.2% of the original root MSE, in both robust methods (Fig. 3). Note that in Finance, "Method 1" is worse than the original estimator. There is also room for improvement of "Method 2": we could derive cutoff levels from the historical data, as is suggested in Kokic and Bell (1994), instead of using only contemporaneous data as we did for this paper.

6. Conclusions.

Downweighting influential observations in small aggregations reduces MSE and gives smoother estimates. In larger domains, treatment of outliers also yields some reduction in MSE, however, one must be careful in setting up the model and corresponding tuning parameters. There is also a field for further work in setting up the cutoff values in large samples, such as, deriving the values from State/Industry/month specific historical data.

The approach based on the influence function, considered here for the estimate of employment, may also be extended to other items in the CES, whose estimators have more complicated form, e.g., estimators of the average weekly hours and average hourly earnings.

References:

Esposito, R., J.K. Fox, D. Lin and K. Tidemann (1994). ARIES: A Visual Path in the Investigation of Statistical Data. *Journal of Computational and Graphical Statistics* 3, 113-125.

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79-87.

Kokic, P. N. and P.A. Bell (1994). Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, 10, 419-435.

McConnell, Sheila., and Goodman, William, (1999) ""Recognition of More Than One Possible Trend in Time Series: Redesigned Screening of Microdata in the Current Employment Survey," Proceedings of the Section on Government Statistics, American Statistical Association.

Pfeffermann, D. and Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. Chapter 12 in R. L. Chambers and C. Skinner (eds.), *Analysis of Survey Data,* Chichester: Wiley , pp. 175 - 195

Zaslavsky, A.M., N. Schenker and T.R. Belin (2001). Downweighting influential clusters in surveys: Application to the 1990 post enumeration survey. *Journal of the American Statistical Association*, 96, 858-869.

Industry	#Domains	Original RMSE	C=2, one-step	C=2, iterated
Overall	1911	1835.8	74.1%	70.4%
Mining	6	3653.8	63.7%	49.1%
Construction	153	2125.5	69.5%	64.2%
Manufact, Dur	142	1912.0	66.3%	70.1%
Manufact,Nondur	108	1836.7	68.3%	67.5%
Wholesale	128	1463.5	49.6%	40.7%
Retail	232	910.0	85.9%	89.5%
Transportation	137	2037.6	70.3%	60.4%
Information	58	3112.1	80.8%	77.4%
Finance	217	1416.8	71.0%	69.9%
Business services	178	2216.5	75.1%	60.0%
Education, Health	190	1872.5	72.2%	69.4%
Leisure, Hospitality	220	2050.1	91.4%	94.1%
Other Services	142	1423.8	52.7%	45.4%

Fig.1. Reduction of the original root MSE (in percentage) for September 2003 MSA/Industry level estimates (10 to 50 respondents, March 2002 benchmark)

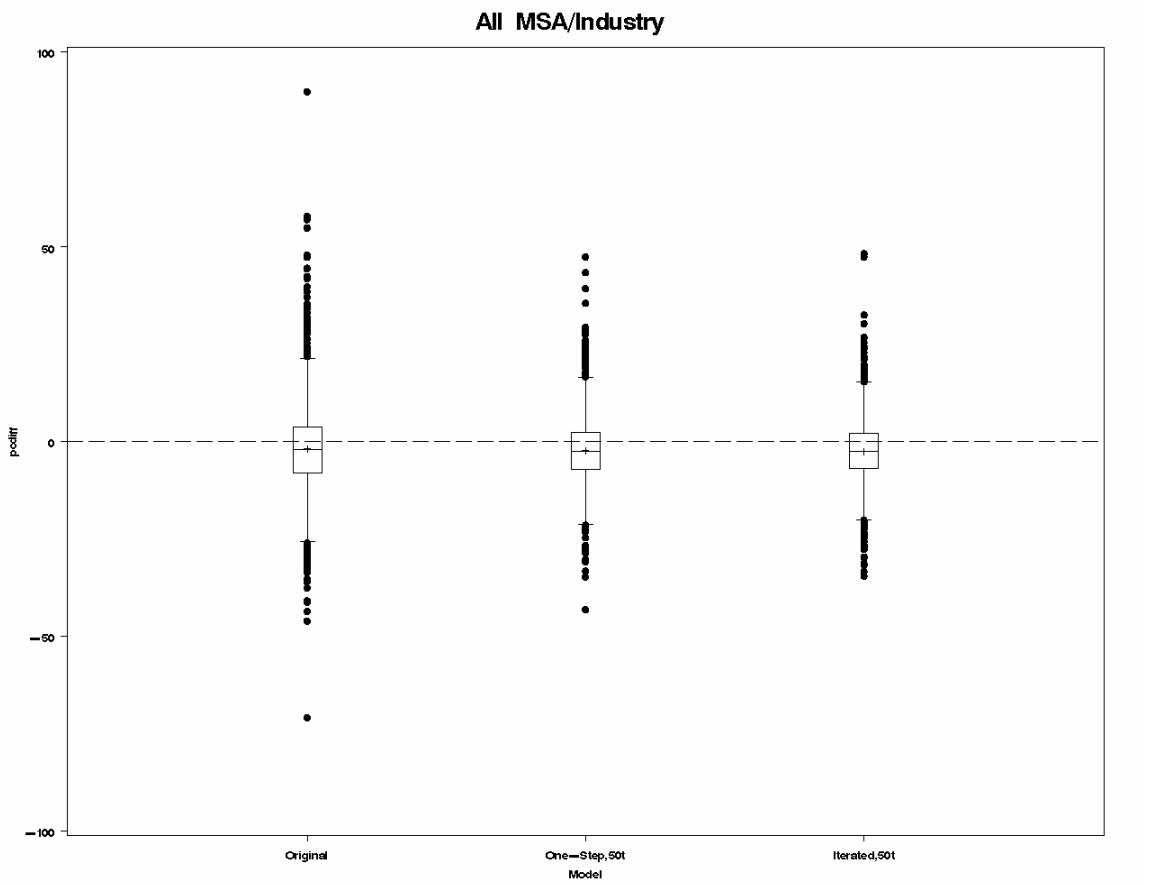

Fig.2 Percent relative error (original estimate, one-step Huber-type estimator with C=2, fully iterated Huber-type estimator

Industry	No. of domains	Original RMSE	Model 1, C=6	Model 2, "Opt L"
Overall	502	12212.9	88.2%	88.2%
Construction	48	6779.7	83.4%	92.1%
Manufact, Dur	42	6452.9	83.6%	63.2%
Manufact,Nondur	35	4368.8	91.1%	92.5%
Wholesale	38	4656.9	74.1%	83.4%
Retail	50	6694.8	95.9%	86.6%
Transportation	37	6801.8	74.7%	91.4%
Information	16	12030.5	89.7%	91.1%
Finance	41	8393.0	111.4%	94.4%
Business services	51	17762.3	86.1%	71.9%
Education, Health	50	17446.8	80.2%	97.2%
Leisure, Hospitality	51	23308.6	91.7%	91.3%
Other Services	43	5400.9	87.2%	94.6%

Fig.3. Reduction of the original root MSE (in percentage) for September 2003 State/Industry level estimates (more than 50 respondents, March 2002 benchmark)

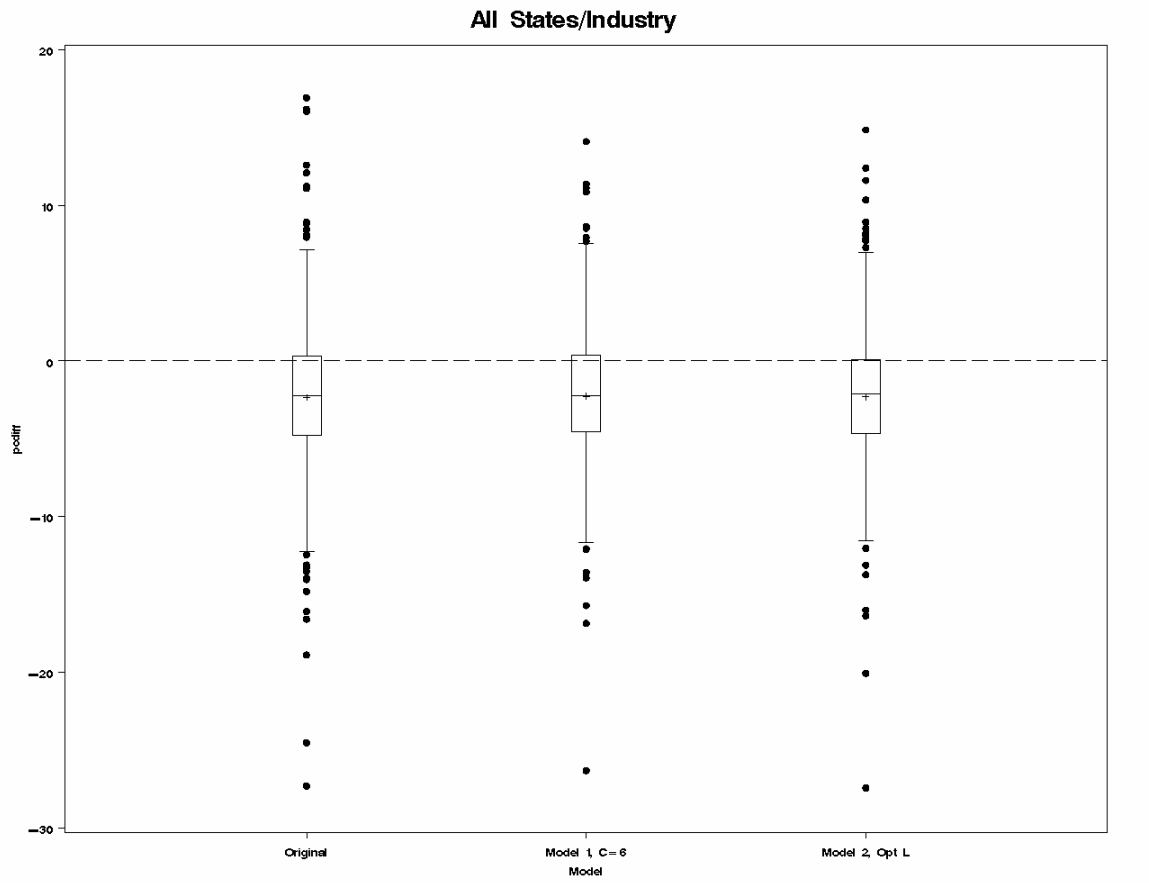

Fig.4 Percent relative error (original estimate, Huber-type estimator with C=6, Model 2 estimator)