Research On Improving The Job Opening and Labor Turnover Survey's Outlier Detection Procedures Using Winsorization Treatment

December 2006

Darrell Greene
Department of Labor, Bureau of Labor Statistics, 2 Massachusetts Ave. NE., Room 4985, Washington, DC 20212

Abstract

The U.S. Bureau of Labor Statistics embarked on the Job Openings and Labor Turnover Survey (JOLTS) in 1999. The JOLTS collects total employment, job openings, hires, quits, layoffs and discharges, other separations, and total separations data. The JOLTS initially used the interquartile method to detect outliers for each characteristic. The list of schedules with potential outliers was reviewed by analysts who determined which schedules were to be treated as atypicals because the value of one or more characteristics was not considered to be representative of the population. There are several statistical issues associated with the initial method. In this paper, we discuss a new procedure to addresses these issues. The new approach uses the Winsorization method.

Keywords: Winsorization method, Interquartile method, outlier

1. Introduction

Starting in 1999, the Bureau of Labor Statistics embarked on the Job Opening and Labor Turnover Survey (JOLTS). A sample of approximately 16,000 business establishments in the private and public sectors, covering all nonagricultural industries for the fifty states and the District of Columbia is selected. The JOLTS collects total employment, job openings, hires, quits, layoffs and discharges, other separations, and total separations data. Before the implementation of JOLTS, there was no economic indicator of labor force demand. The JOLTS serves the purpose of a demand–side indicator.

2. Statement of Problem

The current outlier detection method uses an interquartile approach for each characteristic. The list of schedules with potential outliers is reviewed by analysts who determine which schedules are to be treated as atypicals because value of one or more characteristic is not considered to be representative of the population. At present, the weight of these atypical schedules is changed from their sampling weight to 1.000 so that they represent themselves only.

The concerns with the current procedures are:

1. All the characteristics are treated as being atypical even though most of the time the value of only one characteristic is considered to be an outlier.

2. The reweight of 1.000 is too low. A more appropriate weight is somewhere between 1.000 and the sampling weight.

3. The combination of the above two concerns creates a downward bias for all estimates especially for those characteristics not considered atypical.

4. They rely heavily on analysts' judgment to determine the atypical schedules. Additionally, they increase the processing time and potential for errors arising from manual intervention when reassigning weights.

3. Objective of Research

The purpose of this research is to develop a new procedure that addresses the above stated concerns with the existing procedures. The new approach uses Winsorization methodology. The main purpose of this research is to compare the new procedures to the existing procedures. The new procedures trim the reported value of only the variable that is deemed to be "atypical". In contrast, the existing procedures assign a weight of 1.000 to all variables if any of the variables on the schedule are considered to be an outlier.

The paper is outlined as follows: Section 4 introduces the sources of data; Section 5 discusses the current method interquartile approach; Section 6 covers the Winsorization method; Section 7 compares the two methods; Section 8 details pending questions; and Section 9 makes recommendations.

4. Data Sources

The data used to conduct this research is from SAS datasets and outlier files from estimation retabulations (December 2000 – December 2003). The final SAS datasets and outlier files are from January 2004 – June 2004. The variable(s) that were deemed an outlier were changed to missing. JOLTS imputation then was used on the missing value. The imputed values were treated as being the "true" values.

5. Interquartile Method

JOLTS currently uses an interquartile approach to outlier detection. For each cell (industry, size) with an adequate number of respondents (5 or more) a log transformation of the ratio to employment plus one is made to each variable. The 1st quartile, 3rd quartile (Q3), and the interquartile range (IQR) for each cell are calculated. Units are deemed outliers for a variable if the value of its log transformation falls above the 3rd quartile of the cell by a value greater or equal to c (predetermined value) times the interquartile range of the cell.

If log (ratio + 1) >= (Q3+ c IQR) then outlier.*

With the exception of monthly employment change, no lower threshold is considered for outliers because the lowest possible value a JOLTS ratio may have is zero which is also the most commonly reported value for JOLTS variables. An example is provided (see tables 1 and 2).

Note: The units that do not meet the interquartile test then go through a secondary screening. If the unit's rate (ratio) is greater than a predetermined size rate (ratio) (see appendix table 1) then it is classified as an outlier.

Table 1. Interquartile Method

Unit	Rate	Log(ratio+1)	C	Q3	IQR	Size Rate
1	3.5	0.6532	5	0.545	0.021	2.0
2	0.2	0.0792	5	0.025	0.01	2.0
3	0.5488	0.19	5	0.12	0.016	0.20

Table 2. Continuation

Unit	Rate	Log(ratio + 1)	Q3+(c*IQR)	Size Rate
1	3.5	0.6532	0.65	2.0
2	0.2	0.0792	0.075	2.0
3	0.5488	0.19	0.20	0.20

Unit 1 would be considered as an outlier since the log(ratio + 1) is greater than (Q3 + c * IQR) and the rate is greater than size rate.

Unit 2 would not be considered as an outlier because the rate is less than size rate.

Unit 3 would not be considered as an outlier because log (ratio + 1) is less than (Q3 + c * IQR).

This list of outliers is turned over to analysts. The analysts determine which of these units are not representative of the sample population. ***Only the units designated by analysts as non-representative are treated as outliers in JOLTS estimation.*** Outlier treatment consists of changing the sampling weight, non-response adjustment factor (NRAF), and benchmark factor (BMF) of designated schedules with any outliers to 1.000 so that the schedule is totally self-representing.

6. Winsorization Method

The winsorization method is a procedure that replaces the n extreme values with the preset cut-off value. This method is sensitive to the number of outliers, but not to their actual values. We will not use this technique per se; we will apply the preset cut-off ratio values and multiply them by the reported employment to obtain preset cut-off values.

We used two different winsorized cut-off methods in this exercise: one using size and economic indicators (JOLTS variables) (see tables 3 and 4) and the other using size only (see tables 5 and 6).

Table 3. Winsorization by Size and JOLTS (economic indicators)

Unit	Employ-ment	Hires	Separations	Preset Hires Cut-off	Preset Separations Cut-off
01	25	14	5	0.50	0.50
02	107	55	49	0.40	0.45
03	500	155	184	0.30	0.35

We ratio adjust hires and separations using the preset cut-off ratio times the reported employment to determine the cut-off value for each characteristic. The calculated cut-off values are given in table 4 below.

Hires
unit 01: 25 * .50=12.50

unit 02: 107 * .40=42.80

unit 03: 500 * .30=150.00

Separations
unit 01: 25 * .50=12.50

unit 02: 107 * .45=48.15

unit 03: 500 * .35=175.00

Table 4. Winsorization by Size and JOLTS (economic indicators), continuation

Unit	Hires	Separations
01	12.50	12.50
02	42.80	48.15
03	150.00	175.00

Hires are detected as an outlier for unit 01, 02 and 03 and separations are detected as an outlier for unit 02 and 03 since their reported values exceed the calculated values in the table. Therefore, for unit 01, we will use the value of 12.50 for hires the value of 5 for separations; for unit 02, the values 42.80 and 48.15, respectively, for hires and separations; and for unit 03, the values 150 and 175, respectively, for hires and separations. The sample weights would remain the same.

Table 5. Winsorization by Size Only

Unit	Employment	Hires	Separations	Preset Cut-off
01	25	14	26	1.00
02	107	55	49	0.50
03	500	155	184	0.30

The calculated cut-off values are given in the table 6.

Hires
unit 01: 25 * 1.00=25.00

unit 02: 107 * .50=53.50

unit 03: 500 * .30=150.00

Separations
unit 01: 25 * 1.00=25.00

unit 02: 107 * .50=53.50

unit 03: 500 * .30=150.00

Table 6. Winsorization by Size Only, continuation

Units	Hires	Separations
01	25.00	25.00
02	53.50	53.50
03	150.00	150.00

Hires are detected as an outlier for unit 02 and 03 and separations are detected as an outlier for unit 01 and 03 since their reported values exceed the calculated values in the table. Therefore, for unit 01, we will use the value of 14 for hires and the value of 25 for separations; Therefore, for unit 02, the value of 53.50 for hires and the value of 49 for separations; and for unit 03, the value 150 for both hires and separations. The sample weights would remain the same.

7. Results

We compared the expected error of winsorization defined as EX=(x-hat - μ), where μ equals either the reported value or an imputed value when the JOLTS variable is an outlier {winsorized - μ} versus the expected error of the current method defined as EY=(y-hat - μ) {current – μ}. Two summary statistics were used: 1) sum of errors and 2) sum of absolute value of each error. Both winsorization techniques did better than the current method based on the subset of outliers and the estimates (see appendix tables 4 - 7).

Note: The outlier values used in the tables above are unweighted. The estimates used in the tables above are weighted.

8. Pending Questions

In developing the right cut-offs for the individual (JOLTS) variables one could ask: should it be done by size only, by size/industry or by size economic indicator (JOLTS variables)? Using the winsorized technique more data is subjected to outlier treatment than the current method in which the analysts select the units for treatment. How much of the data should be winsorized?

9. Recommendation and Summary

According to the sum of errors and the sum of absolute value of errors, both winsorization methods of size and size plus economic indicator (JOLTS variables) performed better than the current method from outlier sub-population. However, the current method and winsorization by size+JOLTS did not perform as well as winsorization by size only when comparing estimates. Thus I would recommend using winsorization cut-offs by size at the present time. More research needs to be conducted on the preset cut-offs.

Additionally, the new procedure has the following advantages: 1) each characteristic is independently flagged; 2) outliers are automatically adjusted to a value that is more than self-representing but less than

the sample weighted, thus creating less of a bias than the current procedure.

References

Cox., B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.S., and Kott, P.S. (1995), Business Survey Methods, Wiley and Sons

Barnett, V., and Lewis, T. (1984), Outliers in Statistical Data 2nd Ed., New York, Wiley and Sons

David, H.A. (1981), Order Statistics 2nd Ed., New York, London, Sydney, Toronto, Wiley and Sons

Kokic, P.N., and Bell, P.A. (1994), Optimal Winsorizing Cut-offs for a Stratified Finite Population Estimator, Journal of Official Statistics, 419-435

Preston, P., and Mackin, C. (2002), Winsorization for Generalised Regression Estimation, Australian Bureau of Statistics

Lee, Hyunshik, Outliers in Business Surveys, Statistics Canada, 503-526

Appendix

The size rate (ratio) is a predetermined value for outlier detection based on number of employees of an unit.

Table 1: Predetermined Size Rate

Size	Employment	Size Rate
1	0-9	2.00
2	10-49	1.00
3	50-249	0.50
4	250-999	0.30
5	1000-4999	0.20
6	5000+	0.15

The preset cut-off values in table 2 are the same as the size rates use in the current outlier procedures provided by the analysts .

Table 2: Preset Cut-offs by Size only

Size	Employment	Preset Cut-off
1	0-9	2.00
2	10-49	1.00
3	50-249	0.50
4	250-999	0.30
5	1000-4999	0.20
6	5000+	0.15

We created a database of JOLTS reported values for each JOLTS variable (job openings, hires, quits, layoffs and discharges, other separations, and total separations). The database contained all the JOLTS values ever reported in the JOLTS survey for all collected variables. With this data we were able to determine a distribution of values for each JOLTS variable. From these distributions we could determine percentiles. We used the 98th percentile for a given variable as the cut off for that variable (see table 3).

Table 3: Preset Cut-offs by Size and JOLTS (variables)

Size	Employment	Preset Hires Cut-off	Preset Separations Cut-off
1	0-9	0.50	0.75
2	10-49	0.50	0.50
3	50-249	0.40	0.45
4	250-999	0.30	0.35
5	1000-4999	0.25	0.30
6	5000+	0.15	0.30

Sum of Errors from Outliers

Table 4: Estimates are in thousands.

Method	Job Openings	Hires	Quits	Layoffs & Discharges	Other Separations	Total Separations
Current	25	130	7	51	3	42
Winsorized Size Only	7	60	5	18	2	25
Winsorized Size+JOLTS	4	63	2	6	1	8

Sum of Absolute Value of Each Error from Outliers

Table 5: Estimates are in thousands.

Method	Job Openings	Hires	Quits	Layoffs & Discharges	Other Separations	Total Separations
Current	29	131	9	52	5	46
Winsorized Size Only	8	61	5	21	4	30
Winsorized Size+JOLTS	5	63	3	12	3	17

Sum of Errors from Total Estimates

Table 6: Estimates are in thousands.

Method	Job Openings	Hires	Quits	Layoffs & Discharges	Other Separations	Total Separations
Current	-119513	-162404	-85017	-61755	-12079	-158852
Winsorized Size Only	215	-531	161	-2186	36	-1989
Winsorized Size+JOLTS	-4494	-4105	-3240	-16961	-2182	-22383

Sum of Absolute Value of Each Error from Total Estimates

Table 7: Estimates are in thousands.

Method	Job Openings	Hires	Quits	Layoffs & Discharges	Other Separations	Total Separations
Current	119966	162943	85353	61912	12084	159349
Winsorized Size Only	1128	1523	420	2627	234	2467
Winsorized Size+JOLTS	4494	4111	3240	16961	2182	22383