A New Edit Method

Selvaratnam Sridharma

U.S. Census Bureau

Washington, DC 20233

Disclaimer: Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Abstract

This paper will introduce a new ratio edit method. This new method gives more importance to larger units as the well-known Hidiroglou-Berthelot (HB) ratio edit method does. Here we compare the new edit method with HB edit method using some examples, and the Annual Survey of Public Employment and Payroll (ASPEP). The current method we use for ASPEP is the HB edit method and, it either creates too many outliers or misses many true outliers. This paper will show that the new edit method works better than the HB edit method for certain non-symmetric ASPEP survey data in detecting outliers, and for a symmetric data we created. This new edit method can be used for historical ratios to detect outliers.

Key Words: Outlier, Hidiriglou-Berthelot Edit Method

1. Introduction

Data editing is the process of checking and correcting respondent data in surveys. Outliers are observations that appear to be inconsistent with the rest of the data.

To detect potential outliers in a periodic survey data, several ratio edit methods have been used. One of the most popular ratio edit methods is the Hidiriglou-Berthelot Edit Method (HB Edit method). Usually the HB ratio edit method is preferred over most of the other ratio editing methods since the HB edit method gives importance to larger units. The HB edit method works fairly well for symmetric data for detecting true outliers, but it does not work well for non-symmetric data.

We present here a new ratio edit method to detect outliers that works well for symmetric and non-symmetric data.

In this paper, first, we will go over the HB edit method. Then we will present the New Edit method, and then compare it with the HB edit method using some symmetric data we created.

Finally, we use some data for Annual Survey of Public Employment and Payroll (ASPEP) to compare the New Edit method with the HB edit method. The ASPEP provides state and local government data on full-time and part-time employment, part-time hours worked, full-time equivalent employment, and payroll statistics by governmental function. When we use the HB edit method for ASPEP data, we either get too many false positives – falsely detecting an observation as an outlier, or miss too many true outliers. This happens because ASPEP is mostly non-symmetric data.

These comparisons of methods will show that the New Edit method works better than HB edit method for certain ASPEP data, and some symmetric data we created.

2. HB Edit Method

In the HB edit method (Hidiroglou, 1986), we first calculate the ratio R defined by

$$R = \frac{Y(t)}{Y(t-1)} \quad , \quad \text{where}$$

Y(t) = current period value for a variable, and Y(t-1) = previous period value for a variable

Then we apply a symmetry transformation to R which is defined as,

$$S = \begin{cases} R/R_m - 1 & R \geq R_m \\ 1 - R_m/R & 0 < R \leq R_m \end{cases},$$

where R_m is the median of all ratios that have some common characteristics such as same kind-of-business key, and same level of geographical detail. This transformation centers the distribution of ratios about zero.

For a symmetric data in the example given by the table below, this transformation symmetrizes the data about zero, but it is not completely symmetric about zero.

***Table 1**: Transformed values for S for a sample symmetric data*

Y(t-1)	*Y(t)*	*R*	*S*
100,000	5	0.00005	-24,999
10,000	5	0.0005	-2,499
1,000	5	0.005	-249
100	5	0.05	-24
10	5	0.50	-1.5
1	5	5.00	3
5	1	0.2	-5.25
5	10	2	0.6
5	100	20	15
5	1,000	200	159
5	10,000	2,000	1,599
5	100,000	20,000	15,999

Then we apply a size-effect transformation to ***S*** defined as

$$ES = S\,[\max(Y(t), Y(t-1)]^{u}$$

to place more importance on a small change associated with a 'large' unit as opposed to a large change associated with a 'small' unit. The value of u should be between 0 and 1. This can be changed depending on the data. The following table shows the transformed values for the size effect transformation for the data in the above table.

Table 2*: Transformed values for ES for a sample symmetric data*

Y(t-1)	*Y(t)*	*R*	*S*	*ES*
100,000	5	0.00005	-24,999	-2,499,900
10,000	5	0.0005	-2,499	-99,487
1,000	5	0.005	-249	-3,946
100	5	0.05	-24	-151
10	5	0.50	-1.5	-4
1	5	5.00	3	6
5	1	0.2	-5.25	-10
5	10	2	0.6	2
5	100	20	15	95
5	1,000	200	159	2,520
5	10,000	2,000	1,599	63,657
5	100,000	20,000	15,999	1,599,900

Let $\mathbf{ES_{Q1}}$ = first quartile of **ES**'s, $\mathbf{ES_m}$ = median of ES's, and $\mathbf{ES_{Q3}}$ = third quartile of ES's.

First, we compute, $D_{ES}Q1,A = max\{ES_m - ES_{Q1}, |A*ES_m|\}$ and $D_{ES}Q3,A = max\{ES_{Q3} - ES_m, |A*ES_m|\}$.

Hidiroglou and Berthelot suggested to use **A=.05.** This is to avoid problems when $ES_{Q3} - ES_m$ or $ES_m - ES_{Q1}$ is very small.

Outliers are the units that fall outside $(ES_m - C * D_{ES}Q1,A\ ,\ ES_m + C * D_{ES}Q3,A)$, where **C** is a constant that determines the width of interval.

Finding an appropriate **C** may not be easy. Depending on the value of **C**, either we could have too many false positive outliers or we could miss too many true outliers.

3. New Edit Method

First, we calculate the ratios *R* and *RR* defined as, $R = \frac{Y(t)}{Y(t-1)}$, $RR = \frac{Y(t-1)}{Y(t)}$, where Y(t) = current period value for a variable, and Y(t-1) = previous period value for a variable.

Then we apply the following symmetry transformations to R and RR.

$S = \log(R/R_m)$ and $SR = \log(RR_m/RR)$, where R_m is the median (defined as below) of all ratios *R*, and RR_m is the median of all ratios *RR*. Each transformation centers the distribution of ratios *R* and *RR* about zero.

When the number of observations is even, we define median as geometric mean of middle values when values are arranged in increasing order to find R_m and RR_m **.**Geometric mean of two numbers **a** and **b** is $\sqrt{\boldsymbol{ab}}$. When the number of observations is odd, it is defined as the regular median.

Then it can be easily shown that $RR_m = 1/R_m$**.** For a symmetric data in the following example given by the table below, the transformations *S* and *SR* completely symmetrize the data about zero.

Table 3*: Transformed values for S and SR for a sample symmetric data*

Y(t-1)	Y(t)	*R*	*RR*	*S*	*SR*
100,000	5	0.00005	20,000	-9.90	9.90
10,000	5	0.0005	2,000	-7.60	7.60
1,000	5	0.005	200	-5.30	5.30
100	5	0.05	20	-3.00	3.00
10	5	0.5	2	1.61	-1.61
1	5	5	0.2	0.69	-0.69
5	1	0.2	5	-0.69	0.69
5	10	2	0.5	-1.61	1.61
5	100	20	0.05	3.00	-3.00
5	1,000	200	0.005	5.30	-5.30
5	10,000	2,000	0.0005	7.60	-7.60
5	100,000	20,000	0.00005	9.90	-9.90

Now we apply a size-effect transformation to *S* and *SR* like in the HB edit method. The size-effect transformations are defined as,

$ES = (S)\,[\max(Y(t), Y(t-1)]^{u}$ and $ESR = (SR)\,[\max(Y(t), Y(t-1)]^{u}$.

These transformations place more importance on a small change associated with a 'large' unit as opposed to a large change associated with a 'small' unit.

In the following example, we use u=.5. In the table below, Y(t-1) is previous period value, Y(t) is current period value, *R* is the ratio of Y(t) to Y(t-1), *RR* is the ratio of Y(t-1) to Y(t).

***Table 4**: Transformed values for ES and ESR for a sample symmetric data*

Y(t-1)	Y(t)	*R*	*RR*	*S*	*SR*	*ES*	*ESR*
100,000	5	0.00005	20,000	-9.90	9.90	-3,132	3,132
10,000	5	0.0005	2,000	-7.60	7.60	-760	760
1,000	5	0.005	200	-5.30	5.30	-168	168
100	5	0.05	20	-3.00	3.00	-30	30
10	5	0.5	2	1.61	-1.61	4	-4
1	5	5	0.2	0.69	-0.69	2	-2
5	1	0.2	5	-0.69	0.69	-2	2
5	10	2	0.5	-1.61	1.61	-4	4
5	100	20	0.05	3.00	-3.00	30	-30
5	1,000	200	0.005	5.30	-5.30	168	-168
5	10,000	2,000	0.0005	7.60	-7.60	760	-760
5	100,000	20,000	0.00005	9.90	-9.90	3,132	-3,132

It is not a coincidence that ES is negative of ESR in above table. It is always true for any set of data as it is explained below.

$SR = \log (RR/RR_m) = \log (RR) - \log (RR_m) =$

$\log (1/R) - \log (1/R_m) = -\log(R) + \log (R_m) =$

$-\log (R/R_m) = -S.$

But, $ES = (S)\,[\max(Y(t), Y(t-1)]^{u}$ and

$ESR = (SR)\,[\max(Y(t), Y(t-1)]^{u}.$

This implies ESR = -ES.

For any set of data, if we combine values of ES and ESR, we have a completely symmetric data, which is symmetric about the origin.

This suggests we can define a score for outliers as,

SCORE= |ES|.

Larger values of score will give worse outliers. In this new edit method, we do not need any bounds to detect outliers as in HB edit method. We need to come up with only one parameter U. For the example in the above table, score is given in the following table.

***Table 5**: Scores for a sample symmetric data*

Y(t-1)	*Y(t)*	*R*	*S*	*ES*	*SCORE*
100,000	5	0.00005	-9.90	-3,132	3,132
10,000	5	0.0005	-7.60	-760	760
1,000	5	0.005	-5.30	-168	168
100	5	0.05	-3.00	-30	30
10	5	0.50	-0.69	-2	2
1	5	5	1.61	4	4
5	1	0.2	-1.61	-4	4
5	10	2	0.69	2	2
5	100	20	3.00	30	30
5	1,000	200	5.30	168	168
5	10,000	2,000	7.60	760	760
5	100,000	20,000	9.90	3,132	3,132

4. Identifying Influential Units

For the historical ratios, we can use cell contributions of units to reduce number of potential outliers.

Cell Contribution of a unit for Historical Ratios

A cell refers to all units with common characteristics such as same kind-of-business key, same level of geographical detail, and/or same function.

It is defined as $\frac{|Y(t) - Y(t-1)|}{T(t-1)} * 100,$ where

$Y(t)$ = Current period value for a variable for a unit.

$Y(t-1)$ = Prior period value for a variable for a unit.

$T(t-1)$ = total values for a variable in a cell for previous period.

The most influential units for a cell are the units with largest values of cell contribution. Using the cell contribution, we can reduce the number of potential outliers.

The table below gives the cell contribution of units for a sample data.

***Table 6**: Cell contribution a sample symmetric data*

Y(t-1)	Y(t)	SCORE	CONTR
100,000	5	3,202	90
5	100,000	3,202	90
10,000	5	782	9
5	10,000	782	9
1,000	5	175	1
5	1,000	175	1
100	5	32	0
5	100	32	0
5	1	4	0
1	5	4	0
10	5	3	0
5	10	3	0

5. Comparison of New Edit Method with HB edit Method

Example 1

First, we apply the HB edit Method, and then the New Edit Method for the following data. Here Y(t) is the current data, and Y(t-1) is the previous period data. For half the data, Y(t) is 5 times Y(t-1), and for the rest Y(t-1) is 5 times Y(t). This data is symmetric with respect to zero.

Y(t-1)	Y(t)
10,000	50,000
9,900	49,500
9,800	49,000
.	.
.	.
.	.
200	1,000
100	500

Y(t-1)	Y(t)
50,000	10,000
49,500	9,900
49,000	9,800
.	.
.	.
.	.
1,000	200
500	100

For the HB edit method with u=0.5, A =.05, and C =1.4, we get the following outliers. For these outliers Y(t-1) is always greater than Y(t) even though the original data was symmetric.

Table 7: HB Edit Method

Y(t-1)	Y(t)
50,000	10,000
49,500	9,900
49,000	9,800
48,500	9,700
48,000	9,600

For the New Edit Method, the first six worst outliers are given below. They are symmetric like the original data. Clearly, for this data, the New Edit Method works better than the HB Edit Method.

Table 8: New Edit method

Y(t-1)	Y(t)	SCORE
10,000	50,000	574
50,000	10,000	574
9,900	49,500	571
49,500	9,900	571
9,800	49,000	568
49,000	9,800	568

Example 2

We compared the HB Edit Method and the New Edit Method for full time Pay for ASPEP 2014 data with 2015 data for certain category. Ratios of these data are heavily skewed to the right.

For the HB edit, we used u=0.5, A =.05, and C =40. A=.05 is commonly used, and suggested by Hidiroglou and Berthelot. The HB edit created 62 outliers. These are true outliers. Then we created the outliers for the New Edit Method using u=0.5. We compared the 62 outliers crested by the HB edit method, with the first 62 worst outliers (with the highest scores). Out of these outliers, 52 outliers are common to both methods.

The table below gives the 10 outliers the HB method created that are not among the first 62 worst outliers the New Method created. Also, we give the sum of the differences between current values and the previous period values.

Table 9: HB Edit Method

Y(t)	Y(t-1)	Difference
109,783	23,201	86,582
100,987	19,327	81,660
2,571	26,482	23,911
83,170	13,977	69,193
13,030	80,476	67,446
2,390	29,350	26,960
42,866	3,238	39,628
52,606	5,110	47,496
2,033	26,461	24,428
18,937	96,180	77,243
		SUM = 544,547

Source: Annual Survey of Public Employment & Payroll (2014, 2015), U.S. Census Bureau.

The table below gives the 10 outliers the New Edit Method created that are not among the 62 outliers the HB Edit Method created. Also, we give the sum of the differences between current values and the previous period values.

Table 10: New Edit method

Y(t)	Y(t-1)	Difference
4,356,313	6,122,387	1,766,074
1,056,747	590,442	466,305
753,010	329,004	424,006
1,018,861	1,633,088	614,227
2,335,591	1,507,872	827,719
582,114	243,389	338,725
2,058,909	1,235,660	823,249
398,706	142,478	256,228
721,209	1,235,321	514,112
1,805,220	970,475	834,745
		SUM = 6,865,390

Source: Annual Survey of Public Employment & Payroll (2014, 2015), U.S. Census Bureau.

If we compare the sums of the differences in both tables, clearly the New Edit Method is better than the HB edit Method for U = 0.5 because the sum of differences is much larger for New Edit Method than the HB edit Method.

Keeping the value of U fixed as 0.5 for the New Edit method, and changing values of U for the HB method as 0.3 and then 0.7, we got similar results like before.

Conclusion

In our examples, the New Edit method works better than the well-known and heavily used the HB edit method for symmetric and skewed data. The New Edit method uses only one parameter, but the HB edit method uses three parameters. When a quartile is really close to the median the HB edit could create too many outliers, but the New Edit method will not. When we use the HB edit method bounds, analysts need to give equal importance to every potential outlier. But in the New Edit method potential outliers are ranked by scores so analysts can save a huge amount of time prioritizing outliers.

Acknowledgements

The author would sincerely like to thank Terri Windsor, Paul Villena, and Franklin Winters for their helpful comments on this paper.

References

1. Hidiroglou, M. A. and Berthelot, J.-M. “Statistical Editing and Imputation for Periodic Business Surveys”, Survey Methodology, June 1986, Vol. 12, No. 1, pp 73-83: Journal.