

DYNAMIC OUTLIER DETECTION IN PRICE INDEX SURVEYS

Holly Hutton¹

ABSTRACT

The majority of data sets contain observations that do not conform to the structure followed by the rest of the data. These observations, known as outliers, can be found using a multitude of statistical and non-statistical methods. This paper highlights a generalized system built, specifically for price index surveys, which analysts can use to test different outlier detection methods. It also details the underlying theory involved. With the aid of an analytical user interface, analysts can test these statistical outlier detection methods and immediately see results which will help them reach a decision about the method that best suits their survey data.

KEY WORDS: Generalized system, Outlier detection, Price index.

RESUMÉ

La majorité des ensembles de données contiennent des observations qui s'écartent de la distribution suivie par les autres données. Ces observations, les valeurs aberrantes, peuvent être identifiées en utilisant plusieurs méthodes statistiques et non statistiques. Ce document met en valeur l'élaboration d'un système généralisé construit spécifiquement pour les enquêtes sur les indices de prix, afin que les analystes puissent tester différentes méthodes de détection de valeurs aberrantes. La théorie sous-jacente est également présentée. Ainsi, avec l'aide d'une interface analytique, les analystes peuvent vérifier ces méthodes statistiques de détection de données aberrantes et immédiatement voir des résultats qui les aideront à déterminer quelle méthode est la plus appropriée pour leurs données.

MOTS CLÉS: Données aberrantes; indice de prix; logiciel généralisé.

1. INTRODUCTION

Outliers are observations at either extreme (small or large) of a sample which are so far removed from the main body of the data that the appropriateness of including them in the estimates is questionable. Many statistical methods exist to identify outliers so that appropriate actions can be taken to minimize their impact on survey estimates. The term dynamic outlier detection does not refer to a specific method of detecting outliers but is used in this paper to refer to the creation of a generalized system that allows analysts to study their price index survey data in hopes of developing the best statistical outlier detection strategy possible. This includes the decision of whether or not to impose symmetry functions on the data, the choice of outlier detection method with appropriate parameter values, and an acceptable treatment method.

With many price index surveys already in existence at Statistics Canada and new ones in development for implementation in the near future, the creation of generalized software is a great advantage with many benefits. Specifically, it can decrease survey analysis time and also increase the timeliness in transitioning from the survey pilot testing phase to processing and publishing the new series.

Since the original focus of the generalized system is for price index surveys, Section 2 begins by providing a description of price index surveys and how they differ from other surveys. Sections 3 and 4 present general information about data symmetry and outlier detection. Section 5 discusses general outlier treatment strategies. Section 6 examines the software specifics and Section 7 explains the design of the generalized system and how it implements the outlier detection theory to provide results to the analyst. Section 8 discusses the future goals of the system.

2. OVERVIEW OF PRICE INDEX SURVEYS

Price index surveys are designed to measure the change in the price of goods or services, called items, over time. This requires the collection of prices of the same (or very similar) items as well as explicit descriptions of the items being priced from the same respondent. The collection either takes place on a monthly, quarterly or annual basis and these

¹ Holly Hutton, Statistics Canada, R. H. Coats Building 17-Q, 100 Tunneys Pasture Driveway, Ottawa, ON, K1A 0T6, holly.hutton@statcan.gc.ca

points in time represent the different pricing periods of the price index survey. For example, the price description might include a model number for a particular part used in bike repair or the distance and time required for a courier service to complete a contract. This description is important for verifying that the prices collected are for equivalent items over time.

Survey collection officers will try to ensure the constant quality of the individual prices collected by completing a manual and automated review of prices at the time of collection; however, the analysts study the price relatives to ensure that the price index is not increasing or decreasing excessively, without real justification, over each pricing period. The price relative for an item is the ratio of the current period price over the previous period price. The price index can be calculated in various ways. Generally, a low level elemental index is calculated. This low level index could represent a group of similar items or predefined industrial classes. A higher level index may then be calculated from the elemental indexes to create an overall price index. The higher level index usually incorporates an economic weighting structure so that it is representative of each group's economic contribution to the industry being studied. Four commonly-used price index formulas (IMF - CPI, 2004) are the Carli Index which is the arithmetic mean of price relatives, the Jevons Index which is the geometric mean of price relatives, the Harmonic Mean Index which is the inverse average of price relatives, and the Dutot Index which is the sum of the current period prices over the sum of the previous period prices. These price index formulas can be used for either the elemental or higher level indexes depending on the weighting structure used.

In order to maintain an accurate price index over time, analysts look for and treat outliers in the data. Statistical and/or non-statistical outlier detection methods are applied to the price relative data in hopes of finding the most offending values. The focus here is on statistical outlier detection methods suggested for use with price relative data by the International Monetary Fund (IMF-PPI, 2004), K. Thompson (1999) of the Bureau of Labour Statistics and S. Rubin-Bleuer and A. Saidi (2005) of Statistics Canada. Such methods include the Quartile Method, Tukey Algorithm, Resistant Fences Method, and Median Absolute Deviation Method which are discussed in more detail later in this paper. Unfortunately, price relative data is highly skewed since there is little or no change in the majority of prices collected in each pricing period. This makes the application of statistical outlier detection methods more complex since the majority of them are designed for data with normal or symmetric distributions. To overcome the complexity issues associated with skewed data in the outlier detection process, transformation functions and algorithms exist that can help increase the symmetry of the price relative distribution.

Another complexity of price relative data, that isn't seen in all other types of surveys, is that the extreme values identified by the statistical outlier detection methods could be valid price relatives (i.e. the price collected for an item could have increased or decreased by the amount indicated in the price relative). For this reason when analysts decide on an appropriate treatment method for the identified outliers, it will usually incorporate a manual step to indicate if the treatment method should be applied to the given price relative. Generally accepted outlier treatment methods (that will be described in further detail later in the paper) include data dropping, carry forward imputation, arithmetic mean imputation, geometric mean imputation, and random *hot-deck* imputation. An extra step in the treatment of outliers in price index surveys is that once the treatment method has been applied, the current period price generally needs to be imputed to reflect the new price relative value. Once all outlying price relatives have been treated to the satisfaction of the survey analysts, the impact of the outliers can be seen by recalculating the price index using the new price relatives.

3. SYMMETRIC DATA

Symmetric data is important in most outlier detection methods as they were designed around the assumption of the data following a normal distribution. As a result, formulas used to identify outliers are centred around the median of the data with an interval extending in either direction. If the survey data being studied are not symmetric, a great deal of extra analysis must go into determining the proper parameters in each outlier detection formula.

To obtain a symmetric distribution (or at the very least, less skewed data) from non-symmetric data, the data are generally transformed using a power function. Two widely-used functions are the natural logarithm and the square root functions because they have desirable properties with respect to variance. A very large skewness coefficient for the original data distribution may hamper the ability of these two functions to generate symmetric data. The Box-Cox method is available, in this case, to recommend a power function that stabilizes the variance of the residuals and will generally result in a more symmetric distribution of data. The Box-Cox method uses a parameter lambda (λ) to estimate a power function that should result in nearly symmetric data. Typical λ values are between -2 and +2 (Weisberg, 1985). By maximizing the log-

likelihood function of λ within this range of values, the appropriate transformation for the price relative data can be found. The square root and log functions are possible recommendations from the Box-Cox method.

A separate transformation of survey data used to improve the symmetry of the data is the Hidioglou-Berthelot (HB) transformation. This transformation is different from the power transformations defined above in that is specifically recommended for use with ratio data (Hidioglou and Berthelot, 1986). The HB transformation is a direct function of the properties of the data distribution. For each observation i , the HB transformation (IMF-PPI, 2004) is given below:

$$HB_i = \begin{cases} 1 - \frac{Q_2}{r_i} & 0 < r_i < Q_2 \\ \frac{r_i}{Q_2} - 1 & r_i \geq Q_2 \end{cases} \quad (1)$$

where r_i is the observed survey data and Q_2 is the median of the data distribution. The secondary weighting structure employed by the original HB transformation formula has been omitted here as it does not apply to the theory of price indexes (Rais, 2008).

4. STATISTICAL OUTLIER DETECTION METHODS

As mentioned briefly above, statistical outlier detection methods produce an interval to indicate which survey values are less likely to be erroneous. Some of the many reasons for erroneous data could be errors in the data capture process, respondent providing incorrect information on the questionnaire, or applying manual edits inappropriately. Values falling outside of this interval are considered outliers and should be treated to minimize their impact on the survey estimates. It should be noted that survey data values that are within the interval produced by the outlier detection method could be erroneous but it is statistically less likely. These kinds of erroneous values will not be identified as outliers in a statistical method since they follow the distribution of the survey data.

Four commonly used methods of outlier detection are the Quartile Method, Tukey algorithm, Resistant Fences Method, and Median Absolute Deviation. These four methods are described briefly below. For a more detailed account of the relevant theory, see Rais (2008).

4.1 Quartile Method (QM)

Let Q_2 be the median, Q_1 and Q_3 be the 1st and 3rd quartiles respectively. Then, $(Q_2 - c_L \times \max\{Q_2 - Q_1, |aQ_2|\}, Q_2 + c_U \times \max\{Q_3 - Q_2, |aQ_2|\})$ is the tolerance interval for the Quartile Method. The values of c_U , c_L , and a in the above formula refer to user-defined constants. The parameters c_U and c_L are used to set the width of the tolerance interval while a is meant to compress the width of the tolerance interval to account for data that are highly clustered about the median. When the data are highly clustered about the median, if the Quartile Method formula does not include the maximum function that incorporates the aQ_2 term, then fewer outliers will be flagged than actually exist. The tolerance interval would be too wide to account for the smaller distances between data points. Specifically, a is between 0 and 1.

While the Quartile Method does not require the data be symmetric, implementation of the method is easier with symmetric data since extra work is not required to determine a separate values of c_U and c_L . In this case, $c_U = c_L$. Any of the above-mentioned power transformation can be studied in order to obtain a symmetric distribution. The HB transformation is applied solely with the Quartile Method.

4.2 Tukey Algorithm (TA)

The Tukey Algorithm is used with ratio data, and more specifically, price relative data. As such, this section refers specifically to the Tukey tolerance interval used by the Office of National Statistics in the U.K. to identify outliers in their Consumer Price Index. The tolerance interval is defined as $(M_{10} - c(M_{10} - LM), M_{10} + c(UM - M_{10}))$ where M_{10} is

the trimmed mean (5% trimmed from the top and bottom of the distribution) once price relatives of 1 are removed, LM is the arithmetic mean of price relatives below $Q_{2_{10}}$ (the 10% trimmed median), and UM is the arithmetic mean of price relatives above $Q_{2_{10}}$. In the IMF *Producer Price Index Manual* (2004), the value of M_{10} is substituted for $Q_{2_{10}}$ when outlining the Tukey Algorithm. The c constant is the user-defined constant to define the width of the interval and is set at 2.5 by both the ONS and the IMF manuals.

Due to the nature of the Tukey Algorithm, symmetric data is not required and therefore survey data are not transformed prior to implementation of the method.

4.3 Resistant Fences Method (RFM)

The Resistant Fences Method tolerance interval is $(Q_1 - c_L \times \max\{Q_3 - Q_1, |aQ_2|\}, Q_3 + c_U \times \max\{Q_3 - Q_1, |aQ_2|\})$ where Q_1 , Q_2 , Q_3 and c_U , c_L , and a are as defined in Section 2.1.

In general, the Resistant Fences Method performs better with symmetric data. Any transformation formula besides the HB could be used in this case. Increasing the analysis to define separate values for c_U and c_L allows this method to perform well when using non-symmetric data.

4.4 Median Absolute Deviation Method (MAD)

The tolerance interval for the Median Absolute Deviation Method is given by $(Q_2 - c_L \times MAD_{(r_i - Q_2)}, Q_2 + c_U \times MAD_{(r_i - Q_2)})$ where Q_2 is the median, r_i is the observed survey data for the observation i , and $MAD_{(r_i - Q_2)}$ is the median calculated from all absolute values of $r_i - Q_2$. Again, c_U and c_L are the user-defined constants to set the width of the tolerance interval.

As with most of the other statistical outlier detection methods, data symmetry is a definite asset when applying the Median Absolute Deviation Method. However, by adjusting the c_U and c_L constants, the method will perform well with non-symmetric data as well. Of the mentioned transformation formulas, only the HB transformation is not applicable here.

5. TREATMENT OF OUTLIERS

Once outliers have been identified, one must decide on a strategy to treat them in order to reduce their impact on the survey estimates. Different treatment strategies exist to accomplish this task. The survey data could be modified (imputed), removed, or a weighting adjustment could be made to each observation with an extreme value. The five treatment methods described below are methods used to modify or remove the offending values. Weight adjustments are more complex and won't be described in this paper.

Data dropping is used to remove the extreme values from the estimate completely. The carry forward method imputes the extreme data value with historical survey data from the same respondent. This is used primarily in price index surveys where historical data is available for the same respondent but is possibly used in longitudinal surveys where similar characteristics exist. The arithmetic mean imputation method imputes the extreme data value with the arithmetic mean of all non-outlying observations. This process is usually done within groups or strata so that similar respondent characteristics can be taken into consideration when calculating the imputed value. Geometric mean imputation is the imputation of the extreme data value with the geometric mean of all non-outlying observations. Again, this is generally implemented within groups or strata. The random *hot-deck* imputation method simply imputes the extreme data values with non-outlying values at random. This could be implemented within groups or strata or completely at random from the entire dataset of values.

6. SOFTWARE

The main statistical software used at Statistics Canada is SAS (Statistical Analysis Software). There are many useful functions programmed into SAS to make the creation of generalized systems user-friendly. The generalized system for

outlier detection is built in SAS/AF which allows the user to easily define the necessary parameters in a windows-based, *point and click* atmosphere. All the programming code is behind the scenes and the user is not required to know statistical programming language in order to make use of the system.

7. SYSTEM DESIGN

The generalized system for outlier detection is meant to be an analytical tool to help survey analysts study their data and recommend methodological strategies with respect to outlier detection. Specifically, it is designed with price index surveys in mind, but the generalized nature of the system allows it to be used by other surveys using ratio variables in their estimation process. The generalized system has two functionalities. First is a testing environment for analysts to determine the outlier detection method that best suits their data. This is achieved by studying results based on multiple price index formulas, data transformation formulas to achieve symmetry and outlier treatment methods. Second is a production environment that would allow analysts to detect outliers using a single method (likely identified in the testing environment), treat the *extreme* price relatives and create final output that tracks data changes made. In both, results are generated that need to be reviewed by the analysts. All of these results are displayed within the application as part of the regular system flow. Separate software applications are not required to view them.

7.1 Testing Environment

In the testing environment, the analyst is instructed to identify survey specific information that is used in the outlier detection process. The most important variables identified are the current and previous period prices (so that the price relative can be calculated) and the outlier group variable. The price index formulas and outlier detection methods are applied within these outlier groups (i.e. strata or estimation domains). The first step then is for the system to display the values for each price index formula described in Section 2. Since data symmetry and distribution are very important to choosing an outlier detection method, the generalized system automatically transforms the price relative data using the natural logarithm and square root functions as well as the HB transformation. The system does the necessary calculations to recommend the best λ for the Box-Cox method and indicates which transformation has the best symmetry in the majority of outlier groups. Analysts will get a better feel for their data after studying the histograms of the untransformed and transformed price relatives that are shown for each outlier group. The four outlier detection methods described in Section 4 are applied using all data transformations described above, with the exception of the Tukey Algorithm where only untransformed data are used and the HB transformation which is only applied to the Quartile Method. The system is programmed with default values for the c_U , c_L , and a ; however, the analyst has the opportunity to change these in favour of their own preferred values. Once outliers have been identified using the multiple methods, the analyst is given the opportunity to select one of the treatment methods described in Section 5. The analyst is only able to select a single treatment due to the compile time required to implement all treatment methods on the outliers identified. The final outlier detection results are presented in tables within the application. A single table displays the survey data that have been identified as outliers in the four outlier detection methods. A separate table shows the values of the four price index formulas before and after the outliers are treated. Indication of the total number of outliers for each method is given as well. From this, the analyst is able to identify the transformation formula and outlier detection method that best suits their data. This information can be stored for use by the production environment.

7.2 Production Environment

In the production environment, the analyst is instructed to identify the same survey specific information as in the testing environment; however, the system flow is different as it expects the analyst to know which transformation formula and outlier detection method to use. If the analyst has gone through the testing environment application, the production environment is set up to read in a file of recommendations created at the end of the testing environment run. If the analyst did not go through the testing environment, the system directs the analyst through a flow of screens to identify the most appropriate data transformation formula and outlier detection method. Histograms will be displayed to aid the analyst in choosing a transformation formula; as well the system is set up to provide a recommendation from the Box-Cox method. Similarly to the testing environment, the analyst will be allowed to accept the default values for the c_U , c_L , and a constants or input their own. The main difference between the testing and production environments is the treatment of outliers. The testing environment provided a *one size fits all* application of the treatment methods described in Section 5. The production environment requires the analyst to review all outlying observations and determine if it is an erroneous value that should be treated or if it is an acceptable value that should be left untreated. The outlier treatment is only

applied to the observations indicated by the analyst. As well, the analyst will have the ability to manually change the price relative value. Finally, the results are displayed in a table of price index values before and after treatment. The production environment has *edit* and *save* capabilities that will create the audit trail necessary to verify survey processes in the future. This includes saving any SAS datasets that are created during the process.

8. THE FUTURE OF THE SYSTEM

The overall goal for this system is to be an analytical tool that will help analysts study their survey data in a more efficient and time-constructive manner. This generalized module is for outlier detection but the bigger picture is to incorporate this module into a larger, more comprehensive generalized system that will allow complete survey analysis of other methodological aspects of price index surveys. This may include modules on imputation for non-response, high level index estimation and variance estimation. The challenge is to have a strategy in place that can benefit various price index surveys without being too general as to be unusable in actual survey processing.

Secondary to the overall strategy of having a generalized system for all areas of price index surveys, it would be beneficial if the generalized system (with outlier detection as the starting point) could be available to other Statistics Canada surveys collecting continuous data. Price relatives are a basic ratio of variables so it stands to reason that this generalized system could potentially be used by any survey that deals with ratios of continuous variables. The principles underlying outlier detection, applied to the different surveys, would be the same. The specifics of price index calculation may not be a survey-appropriate way to determine the impact of outliers on the non-price index survey estimates, but it is still a valid indicator that can be studied by survey analysts.

AKNOWLEDGEMENTS

The author would like to thank Zdenek Patak for his guidance in the accomplishment of this work. The author would also like to thank Wesley Yung, Steven Thomas, Gayle Keeley, Andrée Girard and Sylvie Gauthier for their insights, which contributed to improving this paper.

REFERENCES

- Hidioglou, M.A. and J.M. Berthelot (1986). "Statistical Editing and Imputation for periodic business surveys". *Survey Methodology*, June 1986, Vol. 12, N1, pp. 73-83.
- International Monetary Fund (IMF), International Labor Organization, Organization for Economic Co-operation and Development, United Nations Economic Commission for Europe, The World Bank (2004). *Consumer Price Index Manual*. Washington D.C.
- International Monetary Fund (IMF), International Labor Organization, Organization for Economic Co-operation and Development, United Nations Economic Commission for Europe, The World Bank (2004). *Producer Price Index Manual: Theory and Practice*. Washington D.C.
- Office of National Statistics (2006). *Consumer Price Indices Technical Manual*. United Kingdom.
- Rais, Saad (2008). "Outlier Detection for the Consumer Price Index". 2008 Statistical Society of Canada proceedings. *To be published*.
- Rubin-Bleuer, Susana and Abdelnasser Saidi (2005). "Detection of Outliers in the Canadian Consumer Price Index". Statistics Canada for the United Nations Statistical Commission.
- Thompson, Katherine Jenny (1999). "Statistical Methods for Developing Ratio Edit Tolerances for Economic Data". *Journal of Official Statistics*, Vol. 15, No. 4, pp. 517-535.
- Weisberg, Sanford (1985). *Applied Linear Regression* (2nd edition). New York: John Wiley & Sons, Inc.