The study on the application of scanner data in the compilation of price index

Yawen, Liu
University of International Business and Economics
Beijing, China, 100029
liuyawen1985@126.com

Abstract

In nowadays，the technology of scanning bar code is widely used in retail terminals. The scanner data which stem from bar code scanning contains the information of price, sales volume, product code, when and where the product has been sold. This kind of data provides effective foundation for the improvement of compilation of basic component index in price index, by changing the situation that the basic component compilation now cannot use sales volume for weighing but can only make use of price collected by investigators. However, there are still a few technical issues on the application of scanner data. The first issue is how to calculate the unit value to aggregate UPCs across different shops or different weeks to reflect the substitution effect to avoid the overestimation of price index. The second issue is how to update data and meanwhile improve comparability when new products are bought in. In the end, this paper discusses the possible measures can be taken in the future to improve the application of scanner data.

Key words: unit value, substitution effect, chain index

1、Background

Promoting economic growth, controlling inflation, reducing unemployment and maintaining balance of international payments has been every country's most important goals for a long time. As one of the key indicators to reflect inflation, CPI particularly has been the concern of the whole society, so the accuracy of the CPI survey data and the accuracy of index estimation will directly affect the government policies to curb price increase. Besides reflecting the degree of inflation, which affects the direction and intensity of nation's macroeconomic control policies (especially monetary policy), CPI is also an important reference to adjust wages of civil servants, and the minimum wage line for employees. In that case, any tiny flaws in the CPI compilation method are likely to lead to serious consequences.

At present all the countries in the world have chose the fixed basket as the practical framework of CPI. China is no exception. CPI involves a large amount of varieties, indicators, survey points, which makes it extremely complex. In another hand, the fixed basket of CPI intends to ensure that longitudinal comparability. In the 19th and 20th centuries, when economic developed relatively slowly, thus the number of types of goods is small and non-essentials has a smaller proportion, the practice of fixed basket is nothing wrong or even be called delicate, which is both easy to understand and easy to operate. But while economic globalization, the expansion of varieties of trade and consumption is unprecedented and new products emerge in an endless stream, in that circumstance, the basket is difficult to fix.

Faced with the problems brought up by the traditional fixed basket methodology,

some scholars began to explore the possibility of the use of scanner data to compile price index. Scanner data were produced in the supermarkets, shopping malls where bar code of products, which ensured that the goods and scan data were corresponding were scanned, recognized and input into the computer data.

Compared to the traditional price data, scanner data has its own data structures and characteristics. The staff of the U.S. Bureau of Labor Statistics (BLS) has been in contact with several private vendors of scanner data since 1993. Bradley(1995) compared different kinds of index forms with scanner data from A.C.Nielsen. Reinsdorf(1996) calculated the substitution index with the coffer scanner data from Dec.1992 to Dec.1994 in Chicago and Washington. Silver and Heravi(2002)studied the quality adjust with British scanner data of washing machine in 1998. Hauman and Leibtag (2009) discussed the price index error form supermarket.

2、Technique issues of applying scanner data to CPI

Although scanner data has significant advantage, there still are some technique issues under discussion, which can be the key factors of the application of scanner data .

(1) unit value

At present, prices of a representative specification are collected basic monthly at several fixed points on in several sales locations, and then arithmetic average of the several priced are calculated as the month average price of the representative specifications, which is denoted as:

$$\overline{p} = \frac{1}{n}\sum_{i=1}^{n} p_i \ . \tag{1}$$

Unit value is average price weighed by quantity of some goods. One kind of methods to calculate the unit value is sales divided by sale quantity。Unit value is often used in scanner data. On the other hand, there is still controversy on whether the unit value should be considered as price in price index as it is not the price of actual transaction. (Triplett, 2003;Richardson, 2003)

Unit value can be used in several circumstances:

1) calculating the average price of several weeks in a month;
2) calculating the average price of the same goods(UPC) sold in several stores;
3) calculating the average price of similar goods.

For either of the aims above, price and sales volume data in the scanner data can be used to calculate the monthly average price, which is the cumulative sales of the representative specifications divided by cumulative sale quantity, which is denoted as:

$$\overline{p} = \frac{1}{\sum q_i}\sum p_i q_i \ . \tag{2}$$

The sum symbols were not marked, because the meaning here depends on the specific data. If it is weekly data summary, then i represents the number of week; if it is detailed for each transaction, then i means the transaction sequence number.

If the traditional CPI way is adopted to calculate the monthly average price, The price ratio of two month of the representative specifications can be expressed as

follows:

$$\hat{P} = \overline{p}_i^1 / \overline{p}_i^0 = \sum_{i=1}^{n} p_i^{\,1} / \sum_{i=1}^{n} p_i^{\,0} \ , \tag{3}$$

where $\hat{P}$ is not the absolute price but the price ratio of two months.

If monthly average price is calculated as the unit value, the price ratio of two month of the representative specifications can be expressed as follows:

$$\hat{P} = \overline{p}_i^1 / \overline{p}_i^0 = \frac{\sum p_i^1 q_i^1 / \sum q_i^1}{\sum p_i^0 q_i^0 / \sum q_i^0} \tag{4}$$

In this way it can be seen as the summary of sales for the same kind of goods in different stores and different time. As to scanner data, the aggregation is necessary considering the amount of data, on the other hand, average price taking quantity into account, in fact, is a weighted price, equation(4) can be rewritten as:

$$\begin{aligned} \hat{P} \ &= \overline{p}_i^1 / \overline{p}_i^0 = \frac{\sum p_i^1 q_i^1 / \sum q_i^1}{\sum p_i^0 q_i^0 / \sum q_i^0} \\ &= \sum \frac{q_i^1}{\sum q_i^1} p_i^1 / \sum \frac{q_i^0}{\sum q_i^0} p_i^0 \\ &= \sum w_i^1 p_i^1 / w_i^0 p_i^0 \end{aligned} \ . \tag{5}$$

Weights vector in the numerator and denominator in equation(5) are not equal, and therefore equation(5) is different from Laspeyres index or Paasche index.

Unit value will help calculate an appropriate price index in different consumer behaviors, especially in the fast moving consumer goods industry, where the price changes more frequently, there are more sale volume, and price change would be a certain impact on purchasing behavior. When consumer is faced to only one kind of goods, and there is no alternative similar product, according to the theory of consumer, consumers can be divided into the following categories:

a. habit purchaser, who is not sensitive to the price, and will purchase the same goods whether what the price is;
b. common shopper, who is sensitive to the price and will purchase the cheaper goods in the case of sufficient information. But the purchase is limited to this consumer period.
c. Inventory shopper, who will not only purchase the cheaper goods for this consumer period but also purchase and inventory for the next consumer period.

Because of the existence of the second and the third consumer,there will be alternative of one goods in different time or different stores, which makes the weights corresponding to the lower price larger and therefore, when consumers make rational choice, the price index calculated in unit value way is inevitably lower.

A dataset of scanner data from supermarket sales is used here to discuss the unit value. This dataset is scanner data from Dominick's(chained stores in the United

States) provided by Kilts center of the University of Chicago Booth School of Business. This dataset contains sales data of analgesics for four weeks, from Sep. 1989 to May 1997. Kilts Center provides data of a dozen food and fast moving consumer goods, we have chosen to use analgesics based on the considerations that the sales of analgesics does not subject to seasonal effects, and there are few promotional activities affecting the price and sales. The dataset involves a total of 641 kinds of products (UPC), 100 shops, constituting a total of 7,339,217 observations. Variables, including sales volume and price of these goods in each store every week, and the following analysis are data aggregated every four weeks. In that case it is not aggregated monthly strictly but 28 days actually. The final aggregate data set is products sales information of 641 kinds of products for 100 'month'.

Take a product sold in four stores in an area for example, the sales of this product rank in the top ten sales in all of the four stores, with sales data in the table below:

Table 1 sales of one product in four stores

Store number	Quantity of base period	Price of base period	Quantity of reporting period	Price of reporting period	Price ratio	Sales ranking in the store
8	31	4.87	64	4.61	0.95	1
18	30	4.87	38	4.87	1	2
73	26	4.87	75	4.50	0.92	7
94	30	4.87	61	4.54	0.93h	2

It can be seen from the table that the price of this product in the four shops in the base period are the same. While in the second month, this product in some shops cut prices, and some shops did not lower prices. The sales of products with lower prices have been significant growth, at least two times more than the sales of the first month. On the other hand, the store with no price reduce (NO.18 shops), has little increase in sales. Due to the different changes of four shops in the price of the second month, the product sales in four shops share change a lot compared to the base period, indicating the existence of alternative among the four shops. In the case of no other product alternative, when the ups and downs of prices of one product in various shops are not the same, consumers tend to other stores to purchase the cheaper one.

In this case, if the average price in four shops were calculated every month and then price ratio of two months is 0.951; if quantity is taken into account, respectively, with average price equaling the sales divided by quantity, and then price ratio is calculated, the result is 0.944. This means that if we do not consider the substitution effect of the shops, we will overestimate the price of more than 0.7 percentage. This is just one product of a region, if they are cumulated, the result may be in greater deviation. Of course, in this case, this analysis based on the assumption that consumers choose cheaper shops to buy between the same kinds of goods in different shops, without taking into account the cost of this choice, i.e. other differences between the shops, such as services, facilities, etc. . Rational consumer will measure these cost and add them to the price to make a second choice.

(2)comparability

One of the criteria of selecting representative specifications is the so-called stability, that is, if we selected a commodity as the representative specifications, we want it to stay in the market in a long period of time, and to remain in the market accounting for a large share. However, according to product life cycle theory, each product will experience entering-market period, maturity period until out of the market, especially the cycle of high-tech products is particularly short. We can imagine, with the upgrading of products, even the representative specifications we selected still stay in the market, they are no longer "representative". Take the high-tech products for example, along with the product life cycle at different stages, the price changes dramatically. Even if he price movements of the representative specifications keep pace with the overall products, its representative will be a sharp recession considering the role of sales. In addition, representative specifications may also be out of market suddenly, resulting in the price chain disconnected.

In example of analgesics, representative specifications have a longer life cycle, but some high-tech products are perishable, even if they occupy a mainstream position in the market, with the passage of time, the share of sales will continue to reduce until they finally go out of market. Thus it can easily be understood, if the representative specifications just drop in sales, we can ignore the precision and calculate the price index. As soon as they go out of market, it is necessary to find new products for alternative, which will cause a great impact on long-term comparability.

In the categories of products becoming obsolete fast, the emergence of new products will continue. If some new models or even new types of products come into market, the department of statistics should include it in the index as soon as possible. If the product is expected to bring larger sales, the need for that is more. The new product is likely to have a different price changes with existing products, in particular while just entering the market, namely the early stages of product life cycle. If the department of statistics does not find new significant products timely or does not include them in the index, there will be deviations.

The scanner data includes each of the all goods sold. Therefore, if compiling price indices based on scanner data, when a new product emerges, it can be included in the calculation of the index in time. In that case, all items of base period and reporting period can be included and the chain index can be used. The principle of chain index is to use two price indexes of adjacent periods to calculate the price changes of this period, and then make accumulation of changes of this period. If P denotes the price index, the price index in period 0,1,2, respectively, can be expressed as 1, $P(p^0, p^1, q^0, q^1)$, $P(p^0, p^1, q^0, q^1)P(p^1, p^2, q^1, q^2)$. Corresponding fixed base calculations only calculates the price of the period t relative to the price of the base period 0, which is $P(p^0, p^t, q^0, q^t)$. In that case, the price index in period 0, 1, 2, respectively, can be expressed as 1, $P(p^0, p^1, q^0, q^1)$, $P(p^0, p^2, q^0, q^2)$。

Generally statistical agencies use Laspeyres formula to calculate the price index, therefore, the chain index in period 2 can be expressed as

$$P_L = \frac{\sum_{i=1}^{n} p_i^1 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} \frac{\sum_{i=1}^{n} p_i^2 q_i^1}{\sum_{i=1}^{n} p_i^1 q_i^1}$$

If scanner data can be used, a chain index can covers more kinds of goods than the fixed base index. If the fixed base index is adopted, only the products sold in the base period and reporting period at the same time can enter the index compilation. With the use of chain index, as long as we have sales data of two adjacent periods, the current chain index can be calculated and added to index compilation.

3. Future improvement of the application of scanner data

1) Scanner data can only cover the price data of goods with barcode, but not cover service industry. Meanwhile, the places to acquire scanner data is a subset of the traditional CPI survey places, and can not cover all the places where CPI survey take place, because places such as bazaars or groceries do not use scanner so that scanner data cannot be acquire. How to combine the price index from scanner data and the price index from traditional CPI survey is a subject to be discussed in the future.

2) In practical, the application of scanner data is still concentrated in a small amount of market research companies. How to establish a mechanism of reporting to realize the process of massive scanner data from the store directly to the statistics department is an important part for government statistical agencies to consider.

References:

[1]Feenstra, R.C. and Shapiro, W.D. (2003) "High frequency substitution and the measurement of price indexes," *Scanner data and price indexes, NBER Studies in income and wealth*. University of Chicago Press, Chicago.

[2]Fenwick,D., Ball,A., Silver, M. and Morgan, P.H. (2003) "Price collection and quality assurance of item sampling in the retail price index: How can scanner data help?" *Scanner data and price indexes, NBER Studies in income and wealth.* University of Chicago Press, Chicago.

[3]Hawkes,W.J. and Piotrowski, F.W. (2003) "Using scanner data to improve the quality of measurement in the consumer price index," *Scanner data and price indexes*, NBER Studies in income and wealth University of Chicago Press,Chicago.

[4]Reinsdorf, M. B. (1999)"Using scanner data to construct CPI basic component indexes," *Journal of Business & Economic Statistics*, 17, 152-160.

[5]Silver, M. and Heravi, S. (2001) "Scanner data and the measurement of inflation,"*The Economic Journal*, 111, 383-404.

[6]Triplett, J.E. (2003) "Using scanner data in consumer price indexes: some neglected conceptual considerations," *Scanner data and price indexes, NBER Studies in income and wealth.* University of Chicago Press, Chicago.

Acknowledge：

This work was supported by National Social Science Project of China(Grant No.11&ZD015) and Young Teachers Research Project of UIBE (Grant No. 12QD10).