Investigating the use of approximate expenditure weights for web scraped data in consumer price indices

Heledd Thomas, Methodology Division, ONS, UK

The ONS are researching the use of alternative data sources to replace the existing manual price collection, with both coverage and cost in mind. The problem with web scraped data is the lack of expenditure information available. This analysis makes use of a scanner data set for a single retailer, with price and quantity information for Toothpaste and Shampoo products in 2012. The product’s page ranking on a web site is assumed to approximate the quantity sold of that product, thus this analysis ranks the available quantities in the scanner data to approximate web page rankings.

**METHOD 1 – Formulae**

**Goal:** find a formula for transforming product page rankings to product level weights that closely align with weights calculated from expenditure shares. Assessed candidates were:

\[ W(1) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{r_i}{n} \right) \]

\[ W(2) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{r_i}{\text{Rank}(i)} \right) \]

\[ W(3) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{r_i}{\text{Rank}(i)^2} \right) \]

Rank Weights 1 and 2 make use of descending order ranks, i.e. rank 1 is assigned to the most popular product, where \( r_i \) is the rank of product \( i \) and \( n \) is the number of unique products.

Rank Weight 3 makes use of ascending order ranks, i.e. rank 1 is assigned to the least popular product and, in Figure 1, \( x = 6 \).

The benchmark is the expenditure-weighted Geometric Laspeyres index. The index series using weights derived from the Rank 3 method is closest to the expenditure-weighted index, so this is the preferred choice.

**METHOD 2 – Distributions**

**Goal:** find a statistical distribution that suitably approximates the observed quantities and use this distribution to predict sales quantities from their ranks, allowing retailers to provide summary statistics instead of the product-level data set.

Sales quantity ranks are translated to quantiles of cumulative distribution of sales:

\[ F(q_i) = 1 - \frac{r_i}{n} \]

**Observed frequency distributions:** both items’ quantities exhibit long tails, with a very small number of products having very large sales quantities.

**Possible candidates:** log-normal, truncated log-normal and Pareto (power-law).

The truncated log-normal distribution provides the best approximation to the quantities of toothpaste and shampoo. Figure 2 shows how the fitted and observed quantities compare. The resulting indices display similar period-to-period movements to the benchmark expenditure-weighted Geometric Laspeyres index but at a consistently higher level (as shown in Figure 4).

The \( R^2 \) value lies between high 80s and low 90s for each month of 2012 and the Mean Absolute Percentage Error (MAPE) is generally in the mid-20s.

Conclusions

Figure 4 indicates that the Rank 3 weighting method performs better than using the distribution weights for toothpaste, with very little between the two index series for shampoo. Both index series are consistently higher than the expenditure-weighted index series.

A key limitation of the Rank 3 weighting method is that it has only been tested on two items: \( x = 6 \) was optimized on the observed sample of quantities, which would not be available in reality.

A key limitation of using distributions to approximate quantities is that data providers would need to supply the ONS with the required parameters, which may not be p e on an ongoing monthly basis. Thus, the impact on goodness-of-fit of using normalized parameter estimates will be explored, as well as the out-of-sample predictive performance of the fitted truncated log-normal distribution.

Further work is being undertaken to inform ONS’s strategy for incorporating web scraped data into the CPI. The conclusions of this research are limited by the caveat that a scanner data set was used.

Heledd.Thomas@ons.gov.uk

16th Meeting of the Ottawa Group on Price Indices, 8-10 May 2019, Rio de Janeiro