Office for National Statistics

Web Scraped Data: Extreme price changes

Joshua Beeson

Overview

In early 2014 the Office for National Statistics (ONS) Big Data team started developing automated tools to collect prices from retailers' websites. The tools read the underlying HTML to identify price information. This is known as web scraping.

This piece of analysis conducted by the ONS Prices Division is based on over 2.2 million price quotes for 11,155 different food, drink and alcoholic beverages products collected between 2 June 2014 and 9 July 2015. The price quotes were scraped from the websites of supermarket retailers Sainsbury's, Tesco and Waitrose. The analysis focuses on which products change in price to the greatest extent as well as which products change price infrequently.

Methodology

Measuring Volatile Prices

To examine volatile prices in the data set, weekly[1] unit prices for each product were calculated[2]. Using this dataset, price relatives[3] were created with a reference period of the week commencing 2 June 2014. A Tukey test was then conducted on the price relatives to identify the extreme highs and extreme lows. Following the Tukey test, price relatives greater than 2 and less 0.45 were examined in closer detail.

A number of the extreme high and low price relatives could be attributed to one off errors. These were removed from the data set as they did not represent volatile prices. For each product left in the data set, the underlying prices were investigated and found to be genuine price changes, usually as a result of a product coming on or off sale. The results are presented in figures 1 & 2.

Measuring Stable Prices

To examine stable prices, the same price relatives were used. By looking at which products had the highest percentage of price relatives with a value of one, we were able to see which products did not change in price over the measured period. Each of the 11,155 products were placed in 1 of

[1] Analysis was also conducted using fortnightly and monthly unit prices
[2] The price of each product was recorded every day for a week before calculating a geometric average.
[3] A price relative is the ratio of the price of a specific product in one period to the price of the same product in some other period.

35 item groupings. Tests were then carried out to assess the percentage of price relatives that were equal to one within each item. The results are presented in figure 3.

Products with volatile prices

Figure 1: Products that increased by the greatest amount from the base price

Product	Base Price (£)	Extreme High (£)	Extreme Low (£)	Per cent increase
Branded Butter Biscuits	0.50	1.29	0.50	158.0
Branded Chocolate Biscuits	0.79	1.79	0.79	126.6
Branded Ginger Biscuits	0.54	1.09	0.49	101.9
Branded Golden Biscuits	0.54	1.09	0.49	101.9
Branded Ginger Biscuits	0.54	1.09	0.49	101.9
Branded Rich Tea Finger Biscuits	0.54	1.09	0.49	101.9
Branded Chocolate Hazelnut Cereal	1.39	2.80	1.39	101.4
Branded Cola Multipack 8 X 330ML Cans	2.19	4.39	2.00	100.5
Branded Diet Cola Multipack 8 X 330ML Cans	2.19	4.39	2.00	100.5
Branded Chardonnay 75CL	4.99	9.99	4.99	100.2
Branded Shriaz 75CL	4.99	9.99	4.99	100.2
Branded Sauvignon Blanc 75CL	4.99	9.99	4.99	100.2
Branded Shiraz 75CL	5.49	10.99	4.99	100.2
Branded Chardonnay 75CL	5.49	10.99	4.99	100.2
Branded Chardonnay 75CL	5.49	10.99	5.49	100.2

Figure 2: Products that decreased by the greatest amount from the base price

Product	Base Price (£)	Extreme High (£)	Extreme Low (£)	Per cent decrease
Branded Cannelloni	1.99	1.99	0.70	64.8
Branded Pappardelle	1.99	1.99	0.70	64.8
Branded Fettuccini	1.99	1.99	0.70	64.8
Branded Raspberry Biscuits	1.49	1.49	0.64	57.1
Branded Vanilla Biscuits	1.49	1.49	0.64	57.1
Branded Chocolate Biscuits	1.49	2.30	0.64	57.1
Cranberry Juice	2.30	2.30	1.00	56.5
Light Cranberry Juice	2.30	2.30	1.00	56.5
Branded Mature Light Cheddar	4.50	4.50	2.00	55.6

Products with stable prices

Figure 3: Products with the least variation in price

Item	Per Cent of Price Relatives equal to 1
Fresh Onions	96.5
Semi Skimmed Milk	86.9
Fruit Juice, Not Orange	84.9
Branded White Sliced Loaf	80.5
Branded Wholemeal Sliced Loaf	74.2