NBER one page update 7/18/2018

Question: how to introduce alternative data into the CPI.

Alternative data is defined here as any data not collected by traditional field collection procedures performed by BLS staff, and includes third party datasets, corporate data, web scraped data, and data retrieved through API's.

Case Study 1:

BLS receives corporately-supplied data from "Corp Y", a major retail pharmacy accounting for approximately ¼ of the prescription drugs index. In 2012 they refused cooperation with the initiation of new samples because it was too time-consuming for pharmacists. We had discussions with them until May 2014, when we negotiated corporately-supplied data. We received data starting in May 2015 and began using the data in the CPI in May 2016. BLS receives unit prices and some metadata for a sample of prescription drugs.

One lesson that CPI has learned in pursuing corporate data is that it requires compromise, often due to confidentiality concerns of corporate offices. Tradeoffs must be made. Specifically, we think averaging across insurance carriers is a significant compromise, as well as across prescription sizes. However, we feel that in this case the indexes are better with the data they were willing to supply than with no data at all from this company.

Alternative data also has a higher chance of data errors. Corporate data is free of charge and is often a secondary priority for whatever unlucky staffer is responsible for assembling it for us. Even with purchased data, we lose control of quality assurance. For example, CPI caught a data error from this respondent that required a notable correction spanning 4 months.

We will not show Corp Y results because of confidentiality concerns.

Case 2:

BLS has collected daily webscraped prices at 100,000 establishments for three grades of gasoline and diesel. Webscraping does not provide us with weights, so we will be calculating Jevons indexes by area following a matched model approach within establishment. Results will be compared to the CPI indexes.

Case 3:

BLS completed preliminary research using Nielsen ScanTrack data from Sep. 2005-Aug 2010 to evaluate the effect of churn (same product attributes but different UPC) and downsizing. Establishments in Nielsen cover approximately 83% of the CPI sample. Unit values were averaged over establishments, time, and geography. BLS mapped UPCs to CPI item categories and calculated indexes for 43 CPI item strata. Churn and downsizing effects were statistically significant (alpha=0.05) 60 to 80% of the time depending on the index formula used, but the differences were very small (all average differences were 0.1 percentage points or less for the index 12 month percent changes). Future research will be conducted to calculate fixed base Tornqvist indexes to measure chain drift and RYGEKS indexes. Cost and benefit analysis will be done to determine if the purchased data costs less than field collection to determine if it can replace field collection.