A Glimpse into the Future CPI

Improving the CPI

Ralph Bradley

Washington Statistical Society

October 28th, 2014

BLS

BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

www.bls.gov

A Glimpse into the Future CPI

Improving the CPI

Ralph Bradley

Washington Statistical Society

October 28, 2014

The New Challenges

- New online purchasing and payment technology
- Manual price selection
  - Respondent burden
  - Labor intensive
  - Relable PPS Sampling depends on respondent
- Proprietary Prices - that respondents are not willing to disclose.
  - College Financial Aid
  - Medical Prices Negotiated by Private Insurers
- High frequency of product replacement. We run the risk of keeping obsolete products and not having new products.

Outline of Talk

- Discuss the history of how we have addressed these challenges
- Discuss what we have learned
  - Electronically delivered Vendor Data
  - Using other Federal surveys produced outside of BLS
  - Scraping websites of online retailers

What we have implemented for the CPI

- Airline Ticket Prices
  - CPI has been using the Sabre System to get airline ticket prices.
  - Many airlines are leaving the Sabre System, so we are replacing Sabre with the airline websites.
- Medicare Physician Prices
  - A program that keys in a CPT codes for a particular area and downloads the price.
  - This site provides the "allowable" price. Sometimes the transaction price will deviate away from the allowable.

20 Year History -Scanner Data - Before Internet

- Washington and Chicago Coffee (1995)
  - A.C. Nielsen Scanner Data for Washington DC and Chicago IL
  - Marshall Reinsdorf (1999) constructs coffee price indexes for Chicago and Washington
  - He is able to construct superlative indexes.
- New York Cereal Indexes (1998)
  - A.C. Nielsen Scanner Data for 3 PSUs in NY
  - Mass merchandisers did not participate.
  - We still had to use CPI data for these missing stores.
  - David Richardson (2003) constructs cereal price indexes for the New York Areas.
- Replacing CE Diary with Homescan
  - Not all purchases were recorded with Homescan.
  - CE would still have to manually sample households even if Homescan was used.

Uses of Other Electronically Delivered Big Data

- MEPS Disease Based Price Indexes (2007)
  - Combined the Medical Expenditures Panel Survey with CPI data to construct disease based price indexes.
  - This is an example of using existing Federal surveys to improve the CPI.
  - Bradley, Velez, Rozental, Cardenas, Ginsburg (2010) and Bradley (2013)
- CORPX - Department Store Sales
  - A major department store offered CPI its prices for apparel, appliances, household goods, etc..
  - Covered a variety of items.
  - The characteristics accompanying the prices were not adequate for a matched model index.
- JD Power New Vehicle Prince Index (Current)
  - 14 million transaction prices with fields for city, model, trim, drive type, etc. Very easy to incorporate new products.
  - We are investigating replacing manual sample with JD Power data.
  - We can generate 38 Area Indexes and an All Area Index with JD Power data.

Uses of Other Electronically Delivered Big Data (Continued)

- MarketScan Medical Claims Data (Current)
  - 2.5 Terabytes of 600 millions medical claim records.
  - Hospital, Pharmaceutical and Outpatient Files
  - Reimbursements with fields for CPT code, zip code, physician specialty, provider id, insurance type, NDC code, etc.
  - This data is used to evaluate CPI sampling methods - adequate representation of characteristics - correct distribution of prices.
  - We have found -1) The geometric mean assumes too much substitution for pharmaceuticals. 2) CPI physician index is not accounting for consolidation. 3) The change in the DRG classification in 2007 coincided with large price increases that could not be captured in a match model index.
- A.C. Nielsen ScanTrack
  - This data was used to determine if the CPI's PPS sampling produces accurate shares.

Web Scraping

- BLS economist, Ted To, scraped book prices from a large online retailer from August 2013 to May 2014.
- He collected between 260,000 to 330,000 quotes per month along with characteristics that allowed the computation of a matched model index.
- In comparison in the same period, CPI collected between 201 to 266 quotes per month.
- In his study, he uses a Pareto functional form to proxy expenditure shares.
- In his study, he computes a variety of price indexes and compares them to the CPI index for books.
- In October 2013, the price indexes computed from the scraped prices has a large positive jump but the CPI index does not have this jump.

What we have learned.

- Every data source that we have investigated or used was not created for CPI purposes. We always need to make adjustments.
  - Except for MEPS which is a federal survey, the data does not come from a random sample.
  - The geography of the data is not the same as the CPI geography.
  - The characteristics that come with the prices is not as detailed as the characteristics that the CPI uses.
- Big Data can only be implemented in small increments.
  - For vehicles JD Power is a good replacement. But it can only be used for cars.
  - While we can scrape a CMS website for Medicare physician service prices, it does not cover non Medicare prices.
- Big Data should not be used just to collect prices but also to diagnose our index methods.
  - MarketScan Medical Claims
  - ScanTrack

What we have learned.

- Often times the vendor delivers all the data on extremely large files.
- JD Power data set is 14.4 Gigabytes for the 2010-2011 delivery. It contains one third of all US vehicle sales.
- MarketScan data contains 2.5 Terabytes. We use a Unix Sun server to process the data. Jobs take days.
- Index calculation is done far more rapidly by splitting the data into area-month-years. Even the RDB is split.

Challenges of Scraped Data

Scraped data is not sampled on a PPS basis. We need to get expenditure shares. Often times, the sales rank is used and converted.

$$\widehat{s}_{it} = \frac{p_{it} F(\acute{Rank}_{it})}{\sum_{j=1}^{n} p_{jt} F(\acute{Rank}_{jt})},$$

$$F(Rank_{it}) = (Rank_{it})^{-\rho}.$$

This method has problems. If we cannot observe expenditures, there is no way to select the best ρ. Even if we could, there would be errors. For example suppose there are 5 goods that are Cobb Douglas:

$$U = q_1^{.35} q_2^{.30} q_3^{.25} q_4^{.20} q_5^{.15}.$$

Then observed shares are 35%,30%,25%,20% and 15%. Let s_{it} be the true shares. The ρ that minimizes

$$\sum_{j=1}^{n} (\widehat{s}_{jt} - s_{jt})^2$$

is $\rho = .54$.

Even the Best Value Fits Poorly

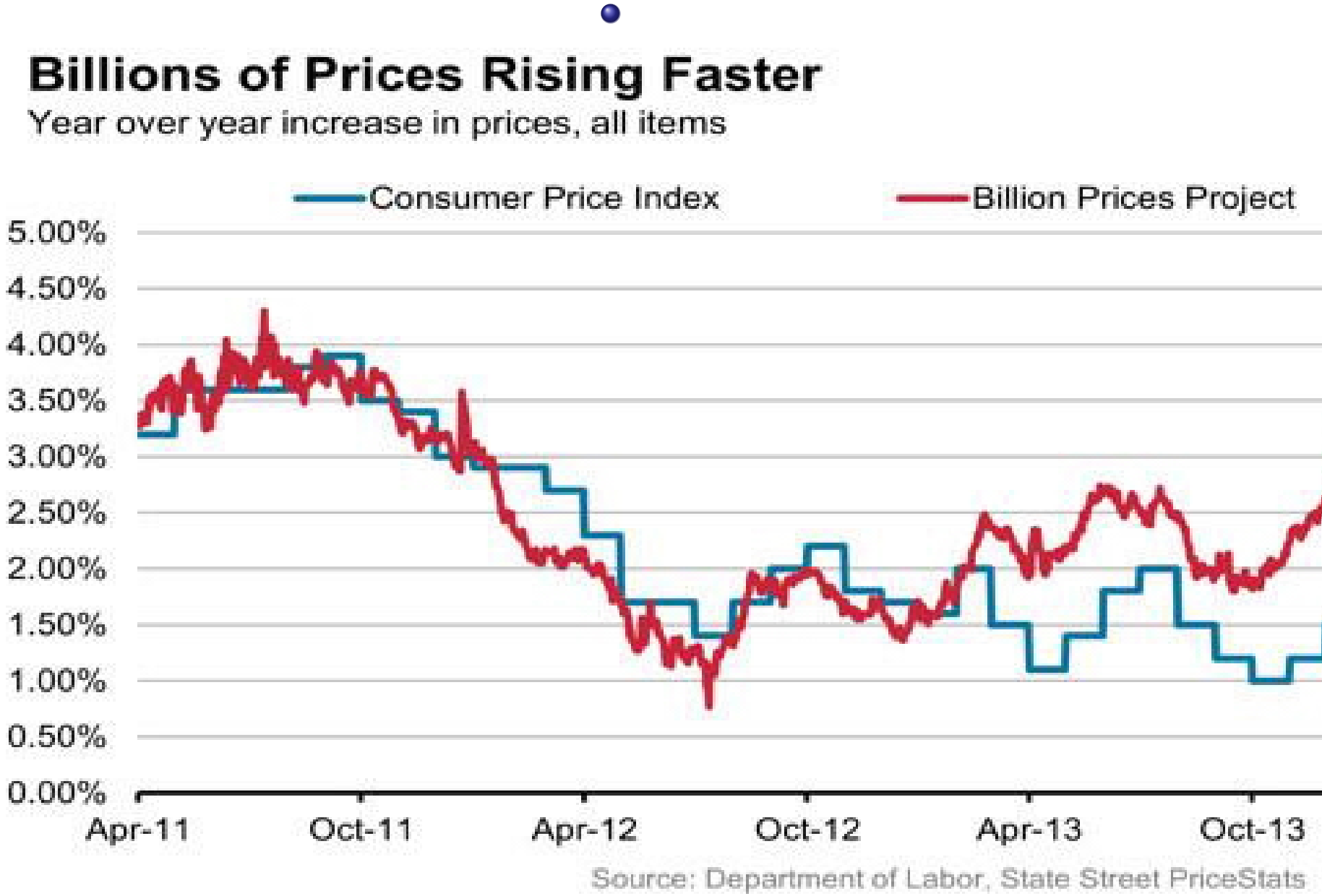
s_{it}	$\widehat{s}_{it}$
.35	.31
.30	.22
.25	.18
.20	.15
.15	.14

- Even the best paramter fits poorly.

Billion Prices Project - Eduardo Cavallo and Roberto Rigobon

- Produces a daily US Price Index through webscraping retailers. CPI is only published monthly.
- Excluded are education, medical, gasoline and other items. No random samples. Weighting methods are not transparent and they do not disclose their weights.
- Has 3,000 CPUs. This is very expensive.
- Scripting programs scrape prices while computing indexes.
  - This is an intense use of highly skilled labor.
  - For large retailers more than one program is required. One for food, one for clothes, etc.
  - The scripts must download the web page source and parse for both prices and characteristics.
  - Often times web pages change so that a program that worked last month will not work this month. The programmer must re inspect the elements containing the prices and characteristics and re write the code.

Comparing the Billion Dollar Price Index with the CPI



What we have learned

- The prices and characteristics on a retailer webpage come from a relational database.
- Running a query on the database is much faster than scraping prices of a dataset. It also is far less labor intensive. Coding does not need to be constantly updated.
- CORPX was willing to query their databases and send us the results. This is a far more efficient transfer of data than web scraping CORPX's websete.
- BLS has been working with big data issues since 1994 before the widespread use of the internet. At that time, no one had anticipated the widespread acceptance of the internet. Perhaps, there is another new technology that will make web scraping obsolete (Json, APIs).

What we have learned

- One area that can improve BLS's price indexes is the use of surveys from other agencies. The diseased based price indexes are such an example. The PPI use of the National Inpatient Sample is another.
- Web scraping can help us identify new goods through high frequency product cycles. It could be an important price source, but we cannot get all our prices through the web.
- We still need to be aware of both the sampling and weighting issues when using an outside data.