Incorporating Alternative Data into the Consumer Price Index: Case Studies

For many years, a primary research objective of the U.S. Consumer Price Index (CPI) has been improving the accuracy of the CPI. More recently the focus has been on the difficult problem of how to introduce alternative data in hard-to-measure areas to more accurately reflect price change, remove quality change, and reduce respondent burden. The purpose of this paper is to present three case studies that reflect our progress on this topic. Case Study 1 focuses on data from a major retail pharmacy, Case Study 2 uses web-scrapped data, and Case Study 3 examines the use of third party datasets.

The Bureau of Labor Statistics (BLS) has traditionally sent Economic Assistants to individual selected establishments around the country to collect price and characteristic data using a tablet computer. In several cases and for a variety of reasons, BLS staff have researched alternative methods to collecting price data. Alternative data is defined as any data not collected by traditional field collection procedures performed by BLS staff, and includes third party datasets, corporate data, and web scraping/API's. Each type of data will be examined in a case study to elaborate on the considerations, and pros and cons of accepting the type of data for the production of the CPI.

Case Study 1: Corporate data:

I will explain our production use of corporate average prices from "CorpY", a major retail pharmacy. Difficult and time consuming in-store procedures for prescription drug sampling and pricing, in addition to privacy concerns at establishments, has led to fragile cooperation for this item across the industry. To reduce respondent burden and increase sample size and therefore accuracy of the index, BLS developed our ideal sampling and pricing methodology for the prescription drug item strata if we can negotiate access to corporate data. I'll explain the ideal, as well as what data we are receiving after negotiation with the company, and problems we've had along the way. Problems include what the corporation is willing to give us versus what we would like to receive, understanding how the data are structured, defined and organized at the company to understand what we can use to sample and identify a unique item to price over time, and the quality of the data and the programs to extract, sample, and average. Results of this establishment's data will be compared to respondents in the CPI, and methods for blending the two data sources will be presented.

Case Study 2: Web scraping:

BLS has begun a webscraping project to collect gasoline prices, with company permission, from a crowd-sourcing application. Gasoline prices are relatively inexpensive to collect, cooperation is not usually an issue with gasoline stations, and there does not appear to be a problem with index accuracy, but with the collection of 48,000 prices in 2017 at about 16,000 establishments, there would be cost efficiencies to be gained by using alternative data. The gasoline webscraping project collects 4 different types of gasoline daily at 100,000 establishments. Problems with webscraping include terms of use, legal issues and continuity of the data source. Webscraping results will be compared to official CPI gasoline indexes, and plans for replacement of the collected data, along with associated shortcomings of using the data, will be discussed.

Case Study 3: Third party datasets:

BLS has received and researched a number of third party datasets. I will present indexes produced using scanner data for Food At Home indexes. Downsizing, churn, frequency of basket updates, and index

formula choice all contribute to the difference between the Nielsen indexes and the CPI. However, these four factors do not on their own or collectively account for all of the disparity. The contributions of downsizing and churn are small on average and are generally overwhelmed by the effect of the index formula and frequency of item updates. The remaining differences between the Nielsen indexes and the CPI could be attributed to several factors including: aggregating UPCs over geographic areas, outlet locations, and outlet types, defining the "item" narrowly at the UPC level (even after accounting for downsizing and churn), geographic and outlet coverage differences, or imputation methods for missing items. Evaluating indexes for scanner data with lower levels of geographic aggregation, ideally only aggregating across UPCs within outlets or within chain/city combinations may provide more insight and be a productive next step toward a scanner data index that could eventually be used in the CPI.