The Effect of CE Sample Sizes on CPI Standard Errors

Jenny Fitzgerald, Harold Gomes, and Imin Hung
U.S. Bureau of Labor Statistics
2 Massachusetts Avenue, NE, Room 3655 Washington, D.C. 20212 U.S.A.
fitzgerald.jenny@bls.gov gomes.harold@bls.gov hung.imin@bls.gov

Abstract

The Consumer Price Index (CPI) estimates the change in prices over time of the goods and services U.S. consumers buy for day-to-day living based on price quotes selected from probability samples and aggregation weights derived from the Consumer Expenditure (CE) Survey. Recently, there has been speculation about the percentage of the CPI-U standard error that is due to the variability of its cost weights. Cost weights are the building blocks of the CPI that are simply the product of CPI indexes and CE aggregation weights. In this study, we draw simulated samples of CE reports to produce simulated cost weights; we then use the simulated cost weights to find a functional form of how the CPI-U all U.S. – all items standard error changes as CE sample sizes change using stratified random groups (SRG) variance methodology.

Key Words: Multistage Sample Design, Simulation Study, Sample Size Variability, Variance Estimation, Stratified Random Groups Methodology

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

1. Introduction

The Consumer Price Index (CPI) estimates the change in prices over time of the goods and services U.S. consumers buy for day-to-day living based on price quotes selected from probability samples along with aggregation weights derived from the Consumer Expenditure (CE) Survey. That is, the Bureau of Labor Statistics (BLS) uses the expenditure data collected by the CE survey to weight the CPI's basic price indexes in order to calculate higher aggregate-level price indexes. The CPI simplifies its aggregate index calculation process by coming up with a set of *cost weights* for each collection period and index series. Cost weights are the product of CPI indexes and CE aggregation weights; they are commonly referred to as the building blocks of the CPI. The BLS sums the CPI basic cost weights over aggregate item groupings (i.e. expenditure classes, major groups, or all items) as well as aggregate areas (i.e. Census regions or the entire U.S.) to calculate aggregate price indexes for each CPI index series. The three main CPI series are: the CPI for all Urban Consumers (CPI-U), the Chained CPI for Urban Consumers (C-CPI-U), and the CPI for Urban Wage Earners and Clerical Workers (CPI-W). Each series has its own set of aggregation weights. Due to the small sample sizes in the basic index areas, the BLS uses composite estimation and a raking smoothing technique to increase the accuracy of the CE aggregation weights produced for each index series.

Recently, there has been speculation about the percentage of the CPI-U standard error that is due to the variability of its cost weights. This paper attempts to answer that precise question. (The BLS only publishes standard errors for its CPI-U index series.) First, brief overviews of the CPI sample design, CPI index estimation process, and CE Survey are provided. Next, the study's methodology is described, in which simulated samples of CE reports are drawn to produce simulated cost weights. Finally, the results of the study are presented, in which standard errors from the simulated cost weights along with their corresponding CE sample sizes are used to find a functional form of how the CPI-U all U.S. – all items standard errors change as CE sample sizes change.

2. CPI Sampling

The CPI-U is calculated from a sample of price quotes, which are the ultimate outcome of several interrelated probability samples. First, the BLS selects a sample of geographic areas, which are the primary sampling units (PSUs) for the CPI (Bureau of Labor Statistics, 2008). The BLS updates its CPI area sample once every ten years. To select its area sample, the BLS divides the entire U.S. into PSUs using the Office of Management and Budget's (OMB) definition of metropolitan statistical areas (MSAs). The BLS then classifies each PSU by its size. A PSU with a population greater than 1.5 million is a self-representing PSU and is given a class size of A. A PSU with a population less than 1.5 million is a non-self-representing PSU. A non-self-representing PSU can be a metropolitan area (with a class size of B) or a non-metropolitan area (with a class size of C).

The second classification variable for PSUs is Census region. After each PSU is mapped to its Census region and given a class-size, the BLS stratifies the PSUs in each region-class size into strata of similar PSUs. Self-representing PSUs are placed in a stratum by themselves; non-self-representing PSUs are stratified based on geographic variables correlated with price change and/or expenditure level. A program then selects one PSU per stratum by means of controlled selection to insure that the selected PSUs are well-distributed across states and to maximize the number of old PSUs selected in the new area sample. Currently, there are 87 PSUs that make up the CPI's 38 index areas.

Within each sampled PSU, the BLS selects a sample of outlets where consumers shop using the data collected via the Telephone Point-of-Purchase Survey (TPOPS) for its Commodities and Services (C&S) sample. TPOPS (which is conducted by the U.S. Census Bureau for the BLS) uses random digit dialling to select a random sample of households. Eligible respondents are asked to provide information about where they bought items and how much they spent during a given recall period for a select group of items (Marsh, 2006). The reported outlets form the frame of outlets that the BLS uses to select its C&S sample for the CPI. The BLS selects its sample of outlets from the frame independently for each PSU, replicate[1], and TPOPS category using a systematic probability proportional to size (PPS) sample design, where each outlet's measure of size (MOS) is its reported expenditure in the TPOPS category.

[1]Each geographic area of the CPI is made of two or more independent samples of items and outlets, called a replicate. A replicate is the basis of the CPI's variance estimates. Independent index estimates are calculated from the replicate samples, while the index produced from the full set of observed prices is called the full sample index estimate. CPI variance estimates are primarily computed using a stratified random groups (SRG) method.

The outlet sample is then merged to an independent sample of ELIs that consumers buy. Specifically, the BLS selects a systematic PPS sample of ELIs for each PSU and replicate combination from the expenditure data collected by the CE Survey, which is aggregated by item stratum and region. An ELI's MOS is its expenditure total for the region compared to the region's total expenditure value for the item stratum. The CPI outlet sample and ELI sample is updated each year for 25 percent of the item strata in each PSU.

Finally, BLS field economists visit the sampled outlets and select individual items for each sampled ELI to be priced each month (or every other month) through a multistage probability sampling technique known as disaggregation for the CPI's C&S sample. The single selection of a unique item is referred to as a price quote (Fuxa, 2010).

For the CPI's housing sample, the BLS ultimately selects housing units in proportion to their share of total rent. First, the BLS defines small geographic areas, called segments, within each of the CPI's 87 PSUs. Segments are one or more Census blocks. Using population and average rent data provided by the U.S. Census Bureau, the BLS estimates the total spent on rent and owners' equivalent rent (OER) in each segment. The BLS then selects a systematic PPS sample of stratified segments where a segment's MOS is its proportion of the total spent on rent and OER in its stratum. Within each sampled segment, field economists select about five renter-occupied housing units to represent the segment.

3. CPI Index Estimation

Each month, the CPI collects price data from its on-cycle sampled C&S quotes and housing units from its 87 PSUs. From the collected price data, the BLS calculates price relatives for all monthly and on-cycle bi-monthly elementary indexes for the CPI. An elementary index is a basic item stratum and basic index area combination. In the CPI, there are 211 basic item strata and 38 basic index areas. Thus, the CPI consists of 8,018 elementary or basic indexes (211×38).

Most elementary indexes use an expenditure-share-weighted geometric average $PRX^{G}_{t,t-1}$ for price relative calculation; other elementary indexes use the Laspeyres formula average $PRX^{L}_{t,t-1}$ (Bureau of Labor Statistics, 2008). The formulas for $PRX^{G}_{t,t-1}$ and $PRX^{L}_{t,t-1}$ are as follows for each index area *a* and item stratum *i* combination:

$$PRX^{G}_{t,t-1} = \prod_{j \in a,i} \left(\frac{P_{j,t}}{P_{j,t-1}} \right)^{\frac{W_{j,POPS}}{\sum_{k \in a,i} W_{k,POPS}}} \quad (1)$$

$$PRX^{L}_{t,t-1} = \frac{\sum_{j \in a,i} \left(\frac{W_{j,POPS}}{P_{j,POPS}} \right) P_{j,t}}{\sum_{j \in a,i} \left(\frac{W_{j,POPS}}{P_{j,POPS}} \right) P_{j,t-1}} \quad (2)$$

Where:

$P_{j,t}$ = the price of the jth observed item in month t for area - item combination a,i;
$P_{j,t-1}$ = the price of the jth observed item in month $t-1$ for area - item combination a,i;
$P_{j,POPS}$ = item j's price in the sampling period when POPS was conducted; and
$W_{j,POPS}$ = item j's weight in POPS.

An elementary full sample (*f*) index value for area *a* and item stratum *i* is calculated by multiplying the previous month's full sample index ($IX_{a,i,f,t-1}$) by the full sample (*f*) price relative for area *a* and item stratum *i* in month *t* ($PRX_{a,i,f,t}$):

$$IX_{a,i,f,t} = IX_{a,i,f,t-1} \times PRX_{a.i,f,t} \quad (3)$$

In the base month (where $t = 0$), the index for area *a* and item stratum *i* is set equal to 100.

$$IX_{a,i,f,t=0} = 100 \quad (4)$$

The CPI item structure has four levels of classification. That is, the CPI's 211 item strata feed into 70 expenditure classes (ECs); the 70 ECs make up eight major groups; and the eight major groups make up the entire CPI. To calculate the aggregated indexes at the EC level and above, elementary indexes are multiplied by aggregation weights derived from tabulated CE data; the product is called a cost weight ($CW_{a,i,f,t}$). The purpose of the aggregation weight is to assign each basic index a relative importance for the higher-level indexes. The cost weights are aggregated to calculate the indexes for the three levels above the elementary index level (which is why cost weights are referred to as the building blocks of the CPI). For example, equation five gives the formula to calculate an index for expenditure class *EC* from area *a* at time *t*:

$$IX_{a,EC,f,t} = IX_{a,EC,f,t-1} \times \frac{\sum_{i \in a,EC} CW_{a,i,f,t}}{\sum_{i \in a,EC} CW_{a,i,f,t-1}} \quad (5)$$

Where:
$IX_{a,EC,f,t}$ = Full sample index for area *a* for expenditure class *EC* at time *t*;
$IX_{a,EC,f,t-1}$ = Full sample index for area *a* for expenditure class *EC* at time $t-1$;
$CW_{a,i,f,t}$ = Full sample cost weight for area *a* - item stratum *i* at time *t*; and
$CW_{a,i,f,t-1}$ = Full sample cost weight for area *a* - item stratum *i* at time $t-1$.

4. The Consumer Expenditure Survey

In order to compute consumer price indexes, the BLS relies on data collected by surveys from programs outside the CPI. One of those surveys is the Consumer Expenditure (CE) Survey. For instance, in Section 2 the expenditure data from the CE Survey serve as an ELI sample frame for the CPI's Commodities and Services quote sample; in Section 3, tabulated CE data function as aggregation weights in CPI index estimation. In the following section, the BLS uses CE expenditures to calculate replicate cost weights for CPI variance estimation. So what is the CE Survey?

The CE Survey actually consists of two surveys: the Quarterly Interview Survey and the Two-Week Diary Survey. Both CE surveys provide information on the buying habits of American consumers by collecting data in extensive detail on the expenditures of each surveyed respondent. The target sample size for both the Interview and Diary Surveys are 7,700 respondents each. The Interview Survey consists of 22 sections, in which respondents describe their expenses from the past three months. Respondents answer the Interview Survey every quarter for a year. The BLS classifies the respondent information reported in the Interview Survey into item categories called universal classification codes (UCCs) (Gamache, 2014). The Diary Survey, in contrast, consists of two consecutive one-week paper diaries. During the reference period, respondents record and describe their expenditures in paper diaries. The BLS later ties the diary respondents' item descriptions to UCCs from the CE program.

So as to use the CE data in the CPI, the BLS maps the UCCs from the Interview and Diary Surveys to the ELIs from the CPI mapping structure. Approximately 600 UCCs map to 311 ELIs. The BLS then adjusts the CE data from both surveys by allocation factors so that the expenditure data meet CPI market basket definitions. CE data also has to be mapped to the geographic areas of the CPI since the two programs geographical areas differ. Specifically, CE geography is based on a 75 PSU design where the population estimates are based on the 2000 Census; the CPI geography, on the other hand, is based on an 87 PSU design (as described in Section 2) where the urban population estimates are based on the 1990 Census. Consequently, the BLS calculates geographic adjustment factors to convert the population estimates for CE areas to population estimates for CPI areas.

Once the CE data is summarized by consumer unit and mapped to the CPI item and geographic structures, the BLS further prepares the expenditure data for use in the CPI's price index and variance estimation process. That is, the BLS expands the price that a consumer unit paid for an item to the estimated cost that all consumers from the same population (i.e. urban, wage earner, or elderly) in the same geographical area paid for that type of item (Casey, 2015). This is done by multiplying the price a consumer unit paid for an item by a population weight, which is then summed to the ELI-UCC-PSU-reference month level for each population group. The BLS then uses composite estimation and a raking technique on the initial monthly cost summaries to derive annual aggregation weights for each CPI item-area combination. The composite estimation technique assigns weights between the basic item - basic area level and that of the basic item – major area level; this process increases the accuracy of the CE aggregation weights by using data from a larger sample at the major area level. The raking technique ensures that the smoothed estimates when summed are equal to the sum of the original estimates at the expenditure class – major area level. The final aggregation weights that the BLS uses to produce CPI cost weights for index estimation and variance estimation are arithmetic averages of two consecutive years of annual weights.

5. Study of the Effect that CE Sample Sizes have on CPI Standard Errors

Recently, there has been speculation about the percentage of the CPI-U standard error that is due to the variability of its cost weights that are derived using CE data. A standard error measures the uncertainty in a survey's estimate caused by the use of a representative sample instead of the complete universe when making a survey's estimates. Sample sizes tend to be inversely related to standard errors. As sample sizes increase, standard errors typically

decrease; and as sample sizes decrease, standard errors usually increase. As stated above, the target sample size for both the CE Interview and Diary Surveys are 7,700 respondents each. If the BLS were to change the sample size of the CE Surveys, how would that modification impact the CPI-U standard errors and by how much?

In January 2015, the BLS charted a team to investigate that very question. The team's research involved drawing simulated samples of CE reports using varying sample sizes to produce simulated cost weights. The team then used the simulated cost weights to calculate standard error estimates for the CPI's all U.S. – all items twelve-month percent change. The team then attempted to use the standard errors and their corresponding CE sample rates to find a functional form of how the all U.S. – all items 12-month CPI-U standard error changes as CE sample sizes change. The research team's methodology is described in detail in the following subsections.

5.1 Sampling of CE Reports

To come up with multiple sets of simulated cost weights, the research team first selected simulated samples of CE Interview and Diary reports of size $c \times S$, where S is the current sample size of the CE Interview and Diary Surveys and c is a constant value or sample rate across all CE PSUs from January 2009 to December 2012. Simulated cost weights require a minimum of two years of data since CE aggregation weights are based on the arithmetic averages of two consecutive years of annual weights. When selecting the samples, the research team pooled the CE data to the CE PSU – data source level (i.e. Diary or Quarterly Interview Survey) for a reference year in order to imitate CE sampling. The team then drew simple random samples (SRS) of CE reports with replacement and equal probability from the pool.

5.2 Derivation of Simulated Cost Weights

With the help of a SAS cost weight simulation program, the research team derived simulated cost weights using the simulated CE samples from Subsection 5.1. As in production, the SAS program expanded the price that a consumer unit (from a simulated sample) paid for an item to the estimated cost that all consumers from the same population and area paid for that type of item. The program then summed the expanded prices to the ELI-UCC-PSU-reference month level, and performed the same composite estimation and raking technique from production on the monthly cost summaries to obtain the annual expenditure weights for each CPI item-area combination for calendar years 2009 – 2012. The final step of the simulation program was to calculate the average expenditures for a two year period for each item-area combination at every level of the CPI item structure[2].

Once the research team had a set of simulated expenditure weights, it could then calculate its simulated cost weights. For both the simulated full sample and replicate cost weights, the team simply multiplied the simulated full sample expenditure weights at the basic index level by the 8,018 basic full sample indexes from production as well as the 20,467 basic replicate indexes[3] from production. The research team then summed the simulated full sample and replicate basic cost weights over aggregate item groupings and aggregate areas to calculate aggregate price indexes all the way up to the all U.S. – all items level. The

[2] The CPI used the expenditure weight data from calendar years 2009 and 2010 and calendar years 2011 and 2012 to calculate its indexes and cost weights from January 2012 – December 2013 and January 2014 – December 2015, respectively.

[3] The BLS calculates CPI replicate indexes using subsets of price relatives from the full sample set of quotes for each area –item combination.

aggregate indexes were then multiplied by the simulated expenditure weights from the aggregate levels of the CPI item structure in order to obtain the simulated cost weights needed for the variance calculations of the all U.S. – all items CPI-U.

5.3 Calculation of Simulated Standard Errors

From the simulated cost weights, the research team calculated twelve-month percent change standard errors for the all U.S. – all items CPI-U. In production, the BLS uses a stratified random groups (SRG) methodology to compute most CPI variances (Shoemaker, 2010). Accordingly, the research team also employed the same SRG variance methodology for its SEs. SRG variance estimation requires a full sample (*f*) percent change (*PC*) and a set of replicate percent changes for area *A* – item *i* at time *t*. Replicate samples consist of two or more independent samples of items and outlets from each CPI index area. The BLS calculates full sample percent changes in the CPI for area *A* – item *i* between months *t* and *t-k* using only full sample cost weights as:

$$PC(A,i,f,t,t-k) = \left(\frac{CW(A,i,f,t)}{CW(A,i,f,t-k)} - 1\right) \times 100\% \quad (6)$$

SRG replicate percent changes, on the other hand, require three sets of data: (1) the full sample cost weights for area *A* – item *i* at months *t* and *t-k*; (2) the set of full sample cost weights for area *a* - item *i,* at months *t* and *t-k* where area *a* is a smaller geographical area within larger geographical area *A* and the complete subset of full sample cost weights for the smaller areas of *A* sum to the full sample cost weight for Area *A* – item *i*; and (3) the complete set of replicate cost weights for area *a* - item *i,* at months *t* and *t-k*. In the CPI, replicate SRG percent changes for area *a*, item *i*, replicate *r*, between months *t* and *t-k* are derived from the full sample and replicate cost weights as:

$$PC(A,i,r,t,t-k) = \left(\frac{CW(A,i,f,t)-CW(a,i,f,t)+CW(a,i,r,t)}{CW(A,i,f,t-k)-CW(a,i,f,t-k)+CW(a,i,r,t-k)} - 1\right) \times 100\% \quad (7)$$

Using formulas six and seven, the research team calculated full sample and replicate twelve-month percent changes for the all U.S. – all items CPI-U from its simulated cost weights from January 2013 – December 2014. Afterward, the team computed its SRG variances as:

$$V[PC(A,i,f,t,t-k)] = \frac{1}{r_a(r_a-1)} \sum_{r=1}^{r_a} (PC(A,i,r,t,t-k) - PC(A,i,f,t,t-k))^2 \quad (8)$$

where:
$A = All\ U.S.(0000)$
$i = All\ items\ (SA0)$
$f = full\ sample$
$t = month\ t$
$k = 12$
$a = the\ 38\ basic\ index\ areas\ of\ the\ CPI$
$r_a = the\ number\ of\ replicates\ for\ basic\ index\ area\ a$

The standard error was then derived as the square root of the variance:

$$SE[PC(A,i,f,t,t-k)] = \sqrt{V[PC(A,i,f,t,t-k)]} \quad (9.)$$

5.4. Accumulation of Multiple Data Points

The process outlined in Subsections 5.1 to 5.3, from the sampling of CE reports to the estimation of the all U.S. – all items CPI-U 12-month standard errors, was repeated for 34 values of *c* (which ranged from 0.25 to 5.0) for ten samples. In total, the research team had 340 points for 24 months to find a functional form of how the 12-month all U.S. – all items CPI-U standard error changes as CE sample sizes change.

6. Results

As a first step in finding a functional form of how the 12-month CPI-U all U.S. – all items standard error changes as CE sample sizes change, the research team applied the Pearson product-moment correlation coefficient (or Pearson correlation coefficient for short) to its data by collection period. The Pearson correlation coefficient (*r*) measures the linear dependence between two continuous variables and is equal to:

$$r = \frac{COV(\text{CE Sample Rate}, \text{CPI 12 - Month Standard Error})}{\sqrt{\text{var}(\text{CE Sample Rate})} \times \sqrt{(\text{var}(\text{CPI 12 - Month Standard Error})}} \quad (10)$$

The value of *r* can range from negative one to one, inclusively, where the sign of *r* indicates the direction of the type of the relationship that exists between two variables and the magnitude indicates the strength of the relationship. Table 1 below gives the Pearson correlation coefficient for the CPI 12-month SEs and CE Sample Rates by collection period.

Table 1. Pearson Correlation Coefficients for the All U.S. – All Items 12-Month CPI-U SEs and CE Sample Rates by Collection Period

Collection Period	Pearson Correlation Coefficient (*r*)
201301	-0.1571
201302	-0.0915
201303	0.0109
201304	-0.0969
201305	-0.1701
201306	-0.1156
201307	-0.0684
201308	-0.1886
201309	-0.0453
201310	-0.0632
201311	-0.0947
201312	-0.1159
201401	-0.2330
201402	-0.3013
201403	-0.2768
201404	-0.1955
201405	-0.3493
201406	-0.1865
201407	-0.1038

Collection Period	Pearson Correlation Coefficient (*r*)
201408	-0.0733
201409	-0.0064
201410	-0.0507
201411	-0.0168
201412	-0.0850

In 23 out of the 24 months, the Pearson correlation coefficient is negative indicating that an inverse relationship exists between the 12-month CPI-U SE and the CE sample rate. The only positive value of *r* appears in March 2013. That value is only 0.0109 though, suggesting that there is no real linear relationship between the two variables. The magnitudes of the other values of *r* are not particularly strong either. The strongest linear relationship between the two variables comes from May 2014, which is -0.3493.

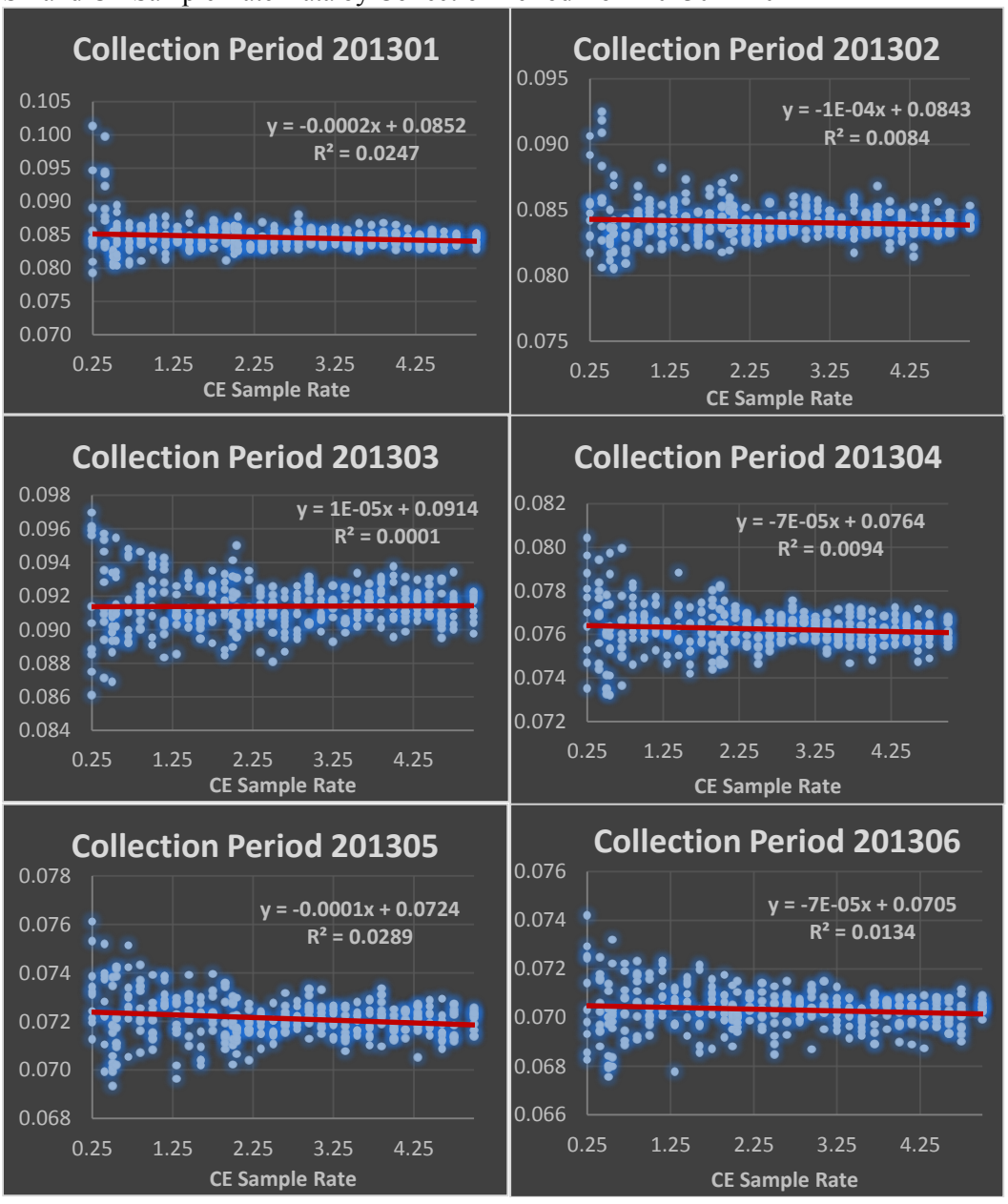
When the research team applied linear regression models to the data by collection period, the team only found results that reinforced the findings from the Pearson correlation coefficient. All of the linear regression models (except for collection period 201303) had negative slopes, but with small magnitudes. Additionally, the r-squared values indicated that the models did not provide adequate explanations of the variance on the 12-month CPI-U standard error. At best, the linear regression model from collection period 201405 explained about 12 percent of the variance of the 12-month all U.S. – all items standard error. Table 2 provides the results of the linear regression models applied to the 12-month CPI-U SEs and CE sample rates by collection period.

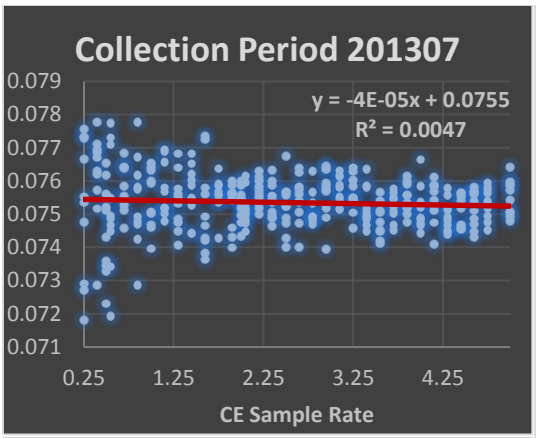
Table 2. Linear Regression Model Summary for All U.S. – All Items 12-Month CPI-U SEs and CE Sample Rates by Collection Period

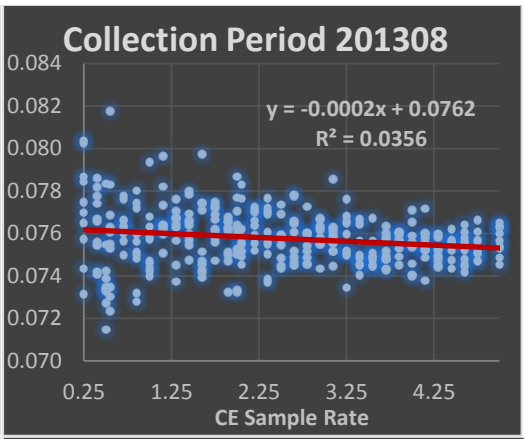
Collection Period	Intercept (β_0)	CE Sample Rate (β_1)	Error Degrees of Freedom	r^2
201301	0.08522	-0.00024	338	0.02467
201302	0.08432	-0.00010	338	0.00836
201303	0.09137	0.00001	338	0.00012
201304	0.07643	-0.00007	338	0.00940
201305	0.07241	-0.00011	338	0.02894
201306	0.07051	-0.00007	338	0.01336
201307	0.07546	-0.00004	338	0.00467
201308	0.07621	-0.00018	338	0.03557
201309	0.07499	-0.00009	338	0.00205
201310	0.08369	-0.00011	338	0.00399
201311	0.07464	-0.00012	338	0.00897
201312	0.07979	-0.00016	338	0.01343
201401	0.07605	-0.00023	338	0.05429
201402	0.07694	-0.00027	338	0.09075
201403	0.07754	-0.00043	338	0.07659
201404	0.06650	-0.00034	338	0.03822
201405	0.07164	-0.00039	338	0.12199
201406	0.08210	-0.00022	338	0.03477
201407	0.08182	-0.00016	338	0.01077

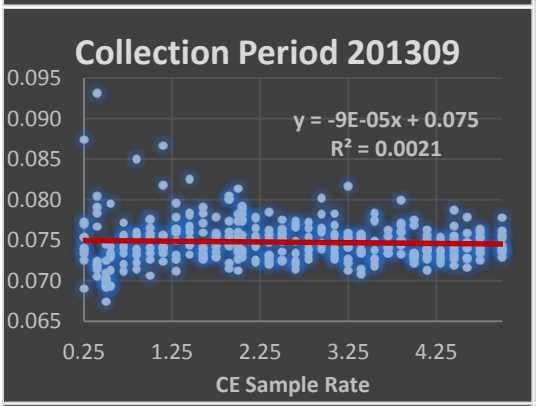
Collection Period	Intercept (β_0)	CE Sample Rate (β_1)	Error Degrees of Freedom	r^2
201408	0.07736	-0.00011	338	0.00537
201409	0.08812	-0.00001	338	0.00004
201410	0.08434	-0.00004	338	0.00257
201411	0.08365	-0.00002	338	0.00028
201412	0.08142	-0.00005	338	0.00723

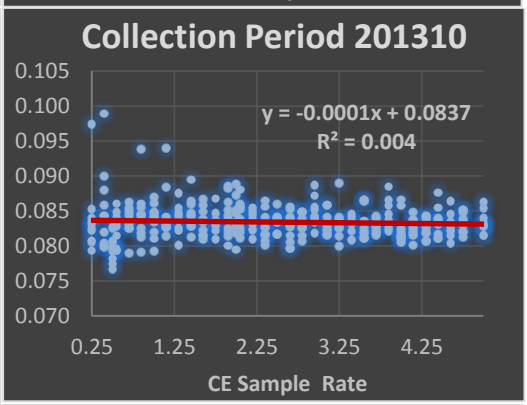
Figure 1. Linear Regression Models Applied to the All U.S. – All Items 12-Month CPI-U SE and CE Sample Rate Data by Collection Period from 201301 - 201412

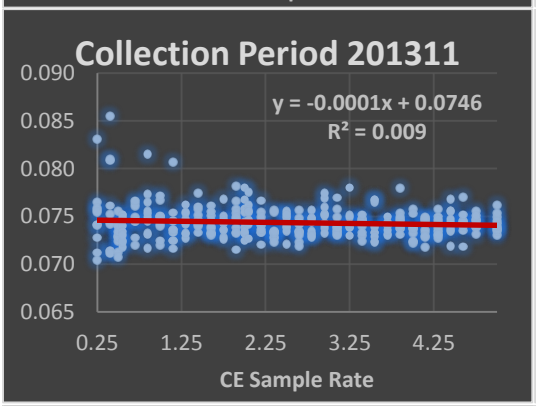


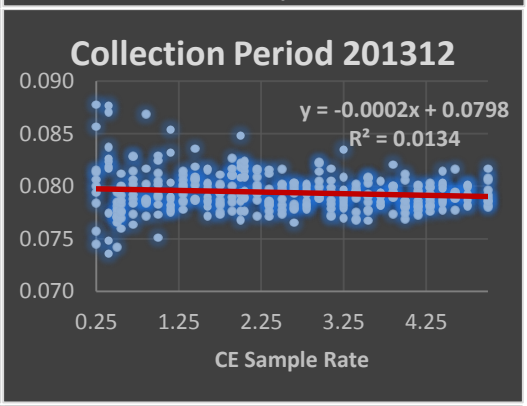

Collection Period 201307
y = -4E-05x + 0.0755
R² = 0.0047
0.079
0.078
0.077
0.076
0.075
0.074
0.073
0.072
0.071
0.25
1.25
2.25
3.25
4.25
CE Sample Rate

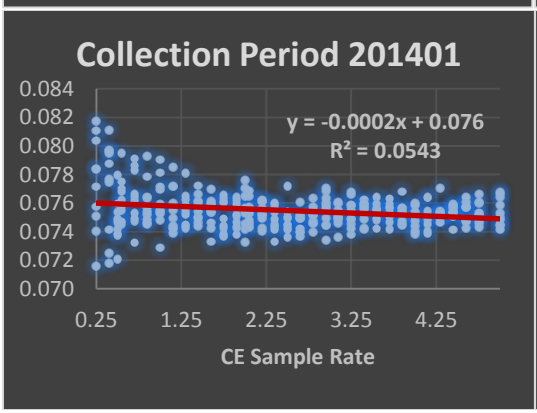

Collection Period 201308
y = -0.0002x + 0.0762
R² = 0.0356
0.084
0.082
0.080
0.078
0.076
0.074
0.072
0.070
0.25
1.25
2.25
3.25
4.25
CE Sample Rate

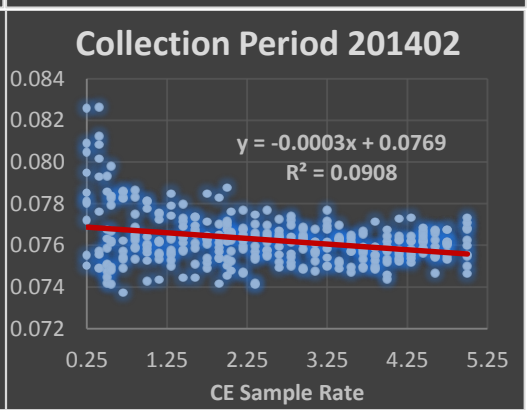

Collection Period 201309
y = -9E-05x + 0.075
R² = 0.0021
0.095
0.090
0.085
0.080
0.075
0.070
0.065
0.25
1.25
2.25
3.25
4.25
CE Sample Rate

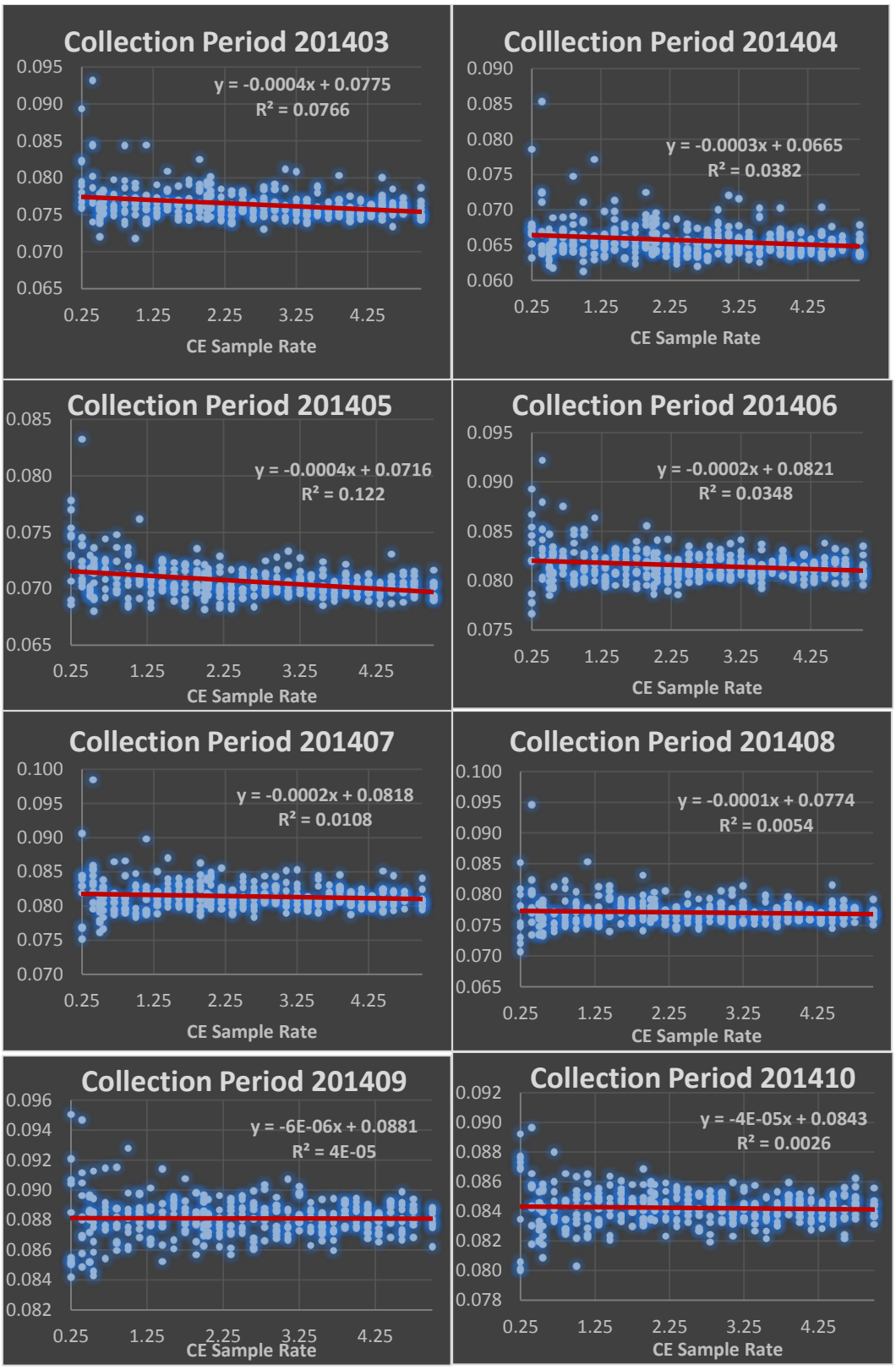

Collection Period 201310
y = -0.0001x + 0.0837
R² = 0.004
0.105
0.100
0.095
0.090
0.085
0.080
0.075
0.070
0.25
1.25
2.25
3.25
4.25
CE Sample Rate

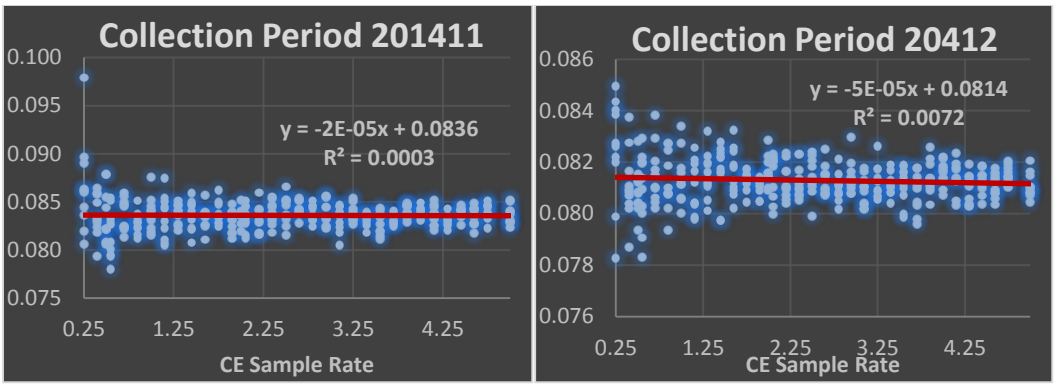

Collection Period 201311
y = -0.0001x + 0.0746
R² = 0.009
0.090
0.085
0.080
0.075
0.070
0.065
0.25
1.25
2.25
3.25
4.25
CE Sample Rate


Collection Period 201312
y = -0.0002x + 0.0798
R² = 0.0134
0.090
0.085
0.080
0.075
0.070
0.25
1.25
2.25
3.25
4.25
CE Sample Rate


Collection Period 201401
y = -0.0002x + 0.076
R² = 0.0543
0.084
0.082
0.080
0.078
0.076
0.074
0.072
0.070
0.25
1.25
2.25
3.25
4.25
CE Sample Rate


Collection Period 201402
y = -0.0003x + 0.0769
R² = 0.0908
0.084
0.082
0.080
0.078
0.076
0.074
0.072
0.25
1.25
2.25
3.25
4.25
5.25
CE Sample Rate


Collection Period 201403
y = -0.0004x + 0.0775
R² = 0.0766
CE Sample Rate
Colllection Period 201404
y = -0.0003x + 0.0665
R² = 0.0382
CE Sample Rate
Collection Period 201405
y = -0.0004x + 0.0716
R² = 0.122
CE Sample Rate
Collection Period 201406
y = -0.0002x + 0.0821
R² = 0.0348
CE Sample Rate
Collection Period 201407
y = -0.0002x + 0.0818
R² = 0.0108
CE Sample Rate
Collection Period 201408
y = -0.0001x + 0.0774
R² = 0.0054
CE Sample Rate
Collection Period 201409
y = -6E-06x + 0.0881
R² = 4E-05
CE Sample Rate
Collection Period 201410
y = -4E-05x + 0.0843
R² = 0.0026
CE Sample Rate



Since linear regression models proved to be poor fits for the data given the weak r-square values, the research team attempted to fit a nonlinear regression model to the data instead of the form:

$$y = c_1 + \left(c_2 \times e^{(-c_3 \times x)}\right) \quad (11)$$

where :

$y = 12\text{-}MonthCPI\text{-}USE\ for\ All\ U.S.\ -\ All\ Items$

$x = CE\ Sample\ Rate$

The team applied the nonlinear regression model to its data in SAS using PROC NLIN. The procedure ran for nine iterations before reaching convergence. Table 3 below provides the parameter estimates for the nonlinear regression model.

Table 3. Nonlinear Regression Model Parameter Estimates for the All U.S. – All Items 12-Month CPI-U SEs and CE Sample Rates

Parameter	Estimate	Approximate Standard Error	95% Confidence Limits
C1	0.0784	0.000073	[0.0783, 0.0786]
C2	0.0081	0.00674	[-0.0051, 0.0213]
C3	6.8712	2.8863	[1.2133, 12.5292]

To evaluate the fit of the nonlinear regression model to the data, a pseudo r-square value was calculated from the residual sum of squares and the corrected sum of squares total of the form:

$$\text{Psuedo } R^2 = 1 - \frac{SS(\text{Residual})}{SS(\text{Total}_{\text{Corrected}})} = 1 - \frac{0.3099}{0.3105} = 0.0019 \quad (12)$$

The pseudo r-squared value of 0.0019 from the nonlinear regression model indicated a poor fit for the data as well. Because the linear and nonlinear regression models failed to provide functional forms explaining how the 12-month CPI-U SE changes as the CE sample size changes, the research team had to further examine its data and consider other possible explanations about the relationship between the two variables.

After reviewing the scatter plot figures from the linear regression models by collection period in Figure 1, what becomes clear is that the spread of the 12-month standard errors

decreases as the CE sample rate increases. Since the research team calculated ten 12-month standard error values for each of the 34 CE sample rate *c* values, the team could then calculate the variance of the standard errors for each sample rate *c*.

Once the research team derived the variance of the 12-month standard errors by sample rate, the team applied the Pearson correlation coefficient (*r*) to its data again (see Table 4 below). Similar to the previous results in Table 1, *r* indicated that a negative relationship exists between the data; but unlike earlier, the Pearson correlation coefficient indicated a much stronger linear dependence between the variables given the greater magnitudes of *r*. To be precise, the Pearson correlation coefficient ranged from -0.4715 in period 201404 to -0.76093 in period 201305.

Table 4. Pearson Correlation Coefficient for Variance of the CPI 12-Month SEs and CE Sample Rates

Collection Period	Pearson Correlation Coefficient
201301	-0.48635
201302	-0.52705
201303	-0.58263
201304	-0.68390
201305	-0.76093
201306	-0.67142
201307	-0.64232
201308	-0.67021
201309	-0.53166
201310	-0.54807
201311	-0.47869
201312	-0.56674
201401	-0.60344
201402	-0.65576
201403	-0.48921
201404	-0.47415
201405	-0.56399
201406	-0.57722
201407	-0.56448
201408	-0.52746
201409	-0.59421
201410	-0.52055
201411	-0.53525
201412	-0.65974

7. Conclusion

Over the past year, there has been speculation within the CPI program about the percentage of its all U.S. – all items 12-month SE that is due to the variability of the program's cost weights. Consequently, the CPI charted a research team to investigate the relationship that exists between those two variables. For its research, the team drew simulated samples of CE Reports to calculate simulated cost weights. From the simulated cost weights, the

research team derived 12-month standard errors that were associated with varying CE sample sizes.

Unfortunately, the team was not able to find a functional form of how the CPI 12-month standard error changes as CE sample sizes change from its samples. The linear and nonlinear regression models were both poor fits for the data. However, what the simulated samples did reveal was that: as the CE sample sizes increase, the variability of the all U.S. – all items 12-month CPI-U standard error decreases. Therefore, larger CE sample sizes will not necessarily reduce the 12-month CPI standard error. On one hand, a smaller CE sample may produce a lower standard error than a larger CE sample. On the other hand, a smaller CE sample can also produce a much larger standard error than the larger CE samples. The spread of the 12-month standard errors from the smaller CE samples were significantly greater than the spread of standard errors from the larger CE sample sizes as demonstrated by the Pearson correlation coefficient.

References

Bureau of Labor Statistics (2008). "Chapter 17 of the BLS Handbook of Methods: Consumer Price Indexes." Accessed February 4, 2015 from the Bureau of Labor Statistics website at http://www.bls.gov/opub/hom/homch17.htm.

Casey, Will, Klick, J., Cage, R., Durgin, Bowman, M., Crosby, M., Paplomatas, A., LE, Rajappa, Larson, W., and M. Saxton (2015). "Use Case Specification: Assemble Initial Expenditure Weights (40)." Internal BLS document dated January 14th, 2015.

Gamache, L., Miller C., and M. Saxton (2014). "Use Case Specification: Process CE Expenditure Data (38)." Internal BLS document dated November 20th, 2014.

Fuxa, Mary Lee (2010). "Sampling and Weighting of Commodity and Service Units for the Elementary Level of Computation of the U.S. Consumer Price Index." *Proceeding Business and Economic Statistics Section, American Statistical Association* [CD ROM], 5014-5025. Alexandria, VA.

Marsh III, Fred (2006). "Comparison between Newly Proposed Response Rates and Current Response Rates for the TPOP Survey." *Proceedings on Consumer Prices and Expenditures Section on Government Statistics, American Statistical Association* [CD ROM]. Alexandria, VA.

Shoemaker, Owen J. (2010). "Variance Estimates for Price Changes in the Consumer Price Index January 2010 – December 2010." Accessed March 20, 2015, 2013 from the Bureau of Labor Statistics website at http://www.bls.gov/cpi/cpivar2010.pdf