Geographic Area Effect in the CPI Variance Model October 2012

Owen J. Shoemaker
U.S. Bureau of Labor Statistics, 2 Mass Ave., NE, Room 3655, Washington, DC 20212
shoemaker_o@bls.gov

Abstract

A long-standing question within the Consumer Price Index (CPI) program has been how best to determine the contribution of geographic areas to the overall CPI variance using standard statistical inference tools. The CPI is constructed of higher-level AREA-ITEM aggregates that are built up from an initial set of AREA-ITEM cells at the basic Index-Area—Item-Stratum level. The CPI produces summary percent price changes for all of these aggregate levels. By utilizing the basic level price changes *and* their higher level price changes, we will proceed to construct an "adaptive" analysis of variance (ANOVA) using these basic level price changes as the initial set of observations. A standard two-way ANOVA with one observation per cell is then applied. The ANOVA results provide F statistics that demonstrate the significance (or not) of AREA and ITEM in the two-way model. For the time periods covered, two out of every three models show AREA to be a significant effect.

Key Words: ANOVA, F test, Components of Variance

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

Introduction

The CPI-U, at its higher (Index) level, is constructed of AREA-ITEM aggregates that build up from an initial set of AREA-ITEM cells at the basic Index-Area—Item-Stratum level. Currently, and since 1997, the number of Index-Areas has been 38. Of these 38 AREAs, 29 of them are A-sized cities (or PSUs), like Denver or Chicago or Miami. Two of them "act" as A-sized cities: Honolulu and Anchorage. These 31 are self-representing AREAs and were selected with certainty in the initial area sample. The other 7 Index-Areas are non-self-representing AREAs and consist of a number of cities (PSUs), each of which represents an optimally sampled set of other cities within the given area stratum, with the chosen city in each stratum representing all the non-chosen cities in that stratum. These 7 non-self-representing Index-Areas include the medium (X-sized) metropolitan and small (D-sized) micropolitan areas throughout the United States. These smaller metropolitan and micropolitan areas, and the geographic strata that they inhabit and represent, are divided into the four natural regions of the country: Northeast (the 100's), Midwest (the 200's), South (the 300's) and West (the 400's). There are 4 X-sized Index-Areas (X100, X200, X300 ,X499), and there are 3 D-sized Index-Areas (D200, D300, D400). Currently there is no D100 because its 1990-based population totals were not large enough to constitute even one full stratum, and its weight and cities were subsumed by X100.

Then within each of these PSUs (or AREAs), the CPI samples and collects monthly and bi-monthly prices in each of 211 Item Strata. There are the larger PSUs (all the A-sized cities), and then there are the smaller PSUs which are the assortment of medium- and small-sized cities in the 7 non-self-representing Index-Areas. In *all* of these PSUs, the CPI currently prices unique items in *all* of these Item Strata on a monthly or bi-monthly basis. The smaller PSUs in the 7 non-self-representing Index-Areas are *sampled* using optimization procedures; the larger PSUs are sampled with certainty and thus are designated self-representing Index-Areas.

In each of the 38 Index-Areas, all of the 211 Item-Strata are sampled on a monthly or bi-monthly basis, producing 38 x 211 = 8018 price relatives each month, which, after updating the previous month's index number in that cell and then being multiplied by its aggregation weight, top and bottom, yields the basic price relative structure: $PREL_t = CW_{a,i,t} / CW_{a,i,t-k}$, where CW = AGGWT * IX, and these CWs are called cost weights. Every higher level price relative, including the one for All-US—All-Items, is simply a sum of the ingredient set of CWs at time *t* in its numerator with the corresponding set of CWs at time *t-k* in the denominator. (NOTE that in the 8018 basic cell price relatives the AGGWTs cancel out, though not in any of the higher level price relatives.) Finally, the price relatives are turned into price *changes* by the simple linear combination: $PC = (PREL_t - 1) * 100$.

An "Adaptive" Analysis of Variance

Analysis of Variance (ANOVA) is the most direct and useful statistical methodology for determining the significance of any one or more effects on the total variance of a model. The question at hand is which ANOVA architecture, if any, is applicable to the structure of price relatives and price changes that we have described in the introduction above. There *seems* to be the outline for a two-way layout with one observation per cell. The two main effects here would be AREA and ITEM, and assuming these two terms to be independent of each other (a strong and proper assumption), there would need to be no interaction terms in the ANOVA model. Moreover, since any higher aggregate CPI estimate is some cross combination of *x* number of AREAs with *y* number of ITEMs, where all *x* AREAs are in every ITEM and all *y* ITEMs are in every AREA, the ingredients for a properly balanced two-way ANOVA seem to be in place. As for the "one observation per cell", this is the basic cell (one AREA by one ITEM) price change. In the ANOVA table below, these price changes are the y_{ij}'s.

So far so good. However, all the *y*-bars in these sums of squares (SS) equations are supposed to be the exact averages of the y_{ij}'s, at the appropriate levels denoted in their subscripts: $y_{\bullet\bullet}$ is the simple average of all the y_{ij}'s; each $y_{i\bullet}$ is the average of each AREA over all the ITEMs in the model; and each $y_{\bullet j}$ is the average of each ITEM over all the AREAs in the model. What we have in the CPI structure are weighted averages in all these $y_{\bullet\bullet}$'s, $y_{i\bullet}$'s, and $y_{\bullet j}$'s, all of which are ratios of averages. ($\Sigma CW_t / \Sigma CW_{t-k}$ is equivalent to $(1/n) \Sigma CW_t / (1/n) \Sigma CW_{t-k}$ since the number of CW's in the denominator will always be the same as the number of CW's in the numerator.) However, *since a ratio of averages is approximately equal to the average of ratios*,

$$(1) \qquad \Sigma^n_i CW_t / \Sigma^n_i CW_{t-k} \approx 1/n \Sigma CW_{t,i} / CW_{t-k,i}$$

Now the LHS of (1) is precisely the price relatives from above and the RHS of (1) is exactly the *y*-bars from the ANOVA table below. (In the RHS all the AGGWTs cancel

out in every computed average, since the weights are always the same for an individual price change at time *t* and time *t-k*.) Moreover, for any least squares calculation any price relative can be translated into a price change without changing the essential ANOVA results. The decimal place simply moves over four places for the Sums of Squares, with the proportions between SSs remaining the same, and more importantly, all the significance test results remain *exactly* the same. Thus, if the CPI higher aggregate —averages" are substituted into the ANOVA table below, then, at least by analogy, *approximately* similar ANOVA results are being produced, as would be the case if the ANOVAs were using their usually prescribed higher-level averages. (In fact, in Table 2, corresponding sets of results will illustrate how similar the two sets of ANOVAs are.)

TABLE 1

Analysis of Variance for Two-Way Layout with One Observation per Cell

Source	SS	d.f.	MS
AREA		I–1	$SS_{Area}/(I–1)$
ITEM		J–1	$SS_{Item}/(J–1)$
ERROR		$\nu_e = (I–1)(J–1)$	SS_e / ν_e
TOTAL		IJ – 1	

Implementing the ANOVA using CPI "averages"

As noted in the introduction, the self-representing Index-Areas and non-self-representing Index-Areas are structurally and stochastically dissimilar parts of the full CPI. The 38 self-representing Index_Areas (the A-Sized Cities) are sampled with certainty while the 7 non-self-representing Index-Areas (the B- and D-Sized Cities) *sample* the set of smaller PSUs that are contained in each of their Index-Areas. The CPI does not calculate summary statistics for the individual PSUs in the B- and D-Sized Index-Areas for any of the Item categories. Thus PSU (or AREA) becomes a random effect in the non-self-representing Index-Areas. Later on, we will look at a set of Variance Components from these non-self-representing Index-Areas to establish at least the percentage influence that AREAs (versus ITEMs or OUTLETs or ERROR) have in this part of the CPI model. For the purposes here, we will look only at the certainty Index-Areas (less Honolulu and Anchorage --- for reasons to be explained shortly) where we have only AREAs and ITEMs, plus ERROR, in the CPI model, and where we have summary (—averages", as it were) statistics for all the variables which are contained in the ANOVA table above. The CPI model will not be All-Cities—All-Items (0000-SA0) but All-A-Sized-Cities—All-Items (A000-SA0). Thus this slightly reduced CPI model will consist of 29 AREAs (all the A's less Honolulu and Alaska) but using all 211 of the ITEMs (or Item-Strata). A000, as a CPI category, itself does not include Honolulu or Anchorage, which is why we will be using A000 (and not A000 + Honolulu + Anchorage) in all the ANOVA calculations.

We need four variables to complete all the Sum of Squares (SS) totals contained in the ANOVA table above: y_{ij}, $\bar{y}_{i.}$, $\bar{y}_{.j}$ and $\bar{y}_{..}$.

- The y_{ij}'s are the 29 x 211 = 6119 basic level 12-month CPI price changes.
- The $\bar{y}_{i.}$'s are the 29 AREA summary 12-month price changes, each with all 211 ITEMs.
- The $\bar{y}_{.j}$'s are the 211 ITEM summary 12-month price changes, each with all 29 AREAs.
- The one $\bar{y}_{..}$ is the All-A-Cities (A000)—All-Items summary 12-month price change.

Using an EXCEL spreadsheet, we then pull in all the 1- and 12-month price changes, from the CPI databases, for the 12 months of 2009 and proceed to calculate all the Sums of Squares (SSs) using the formulas from the ANOVA table above. Knowing that I = 211 and J = 29, we can fill out a complete set of the ANOVA summary tables, including F-test results for the two main effects in the model, AREA and ITEM. We will include a column for Percentage of Total Sums of Squares (% of SS) for each effect and for ERROR. We know the various higher-level summary price changes are not exact averages from the basic cell price changes, so the Total Sums of Squares as calculated will never exactly equal the sum of the three terms in the model (as it has to in any regular ANOVA table). But we will note the percent ratio of that TOTAL to the SUM of the three SSs in the model and use the closeness of that ratio to 100% to gauge how well the "adaptive" methodology here conforms to a true ANOVA structure. Finally, we will add on a Model Standard Error (= Sqrt [MSE/6118]) and compare that to its official CPI SE counterpart. As an added check on the worthiness of the methodology, we will perform a straight-forward ANOVA on the 8,018 basic level 12-month price changes and put them side by side with the "adaptive" ANOVAs for comparison.

ANOVA Results

On the following two pages, the two sets of ANOVA results for the 12 months of 2009 are displayed. The regular ANOVA results are to the right, the "adaptive" ANOVA results to the left. The two sets of results are clearly more similar than not. While the "*y*-bars" in the regular ANOVAs are at no turn equal to or even resemble the CPI price change "means" used in the "adaptive" ANOVAs, the ANOVA results themselves are nearly equivalent at every turn, even the *p*-values from the F tests. The Sums of Squares are roughly equivalent, point by point, in all twelve comparison sets, with the other near equivalencies following naturally from those results. Clearly the "adaptive" ANOVAs are not wildly out of sync with the regular ANOVAs using exact ANOVA methodologies and calculations. This is some degree of evidence that the one main assumption in (1) is a sound enough assumption --- at least as adapted for these ANOVA uses. A second measure of the soundness of the "adaptive" ANOVA is how close to 100% the percent ratio of the calculated TOTAL SS comes to the sum of the three SS terms in the model (i.e, $TOTAL_{SS}$ / ($AREA_{SS}$ + $ITEM_{SS}$ + $ERROR_{SS}$)). The twelve percent ratios average out to 99.23% across the 12 sets of results. Again, more evidence attesting to the soundness of the "adaptive" ANOVA constructions.

TABLE 2

Two-Way ANOVAs for A000-SA0 (LHS = "Adaptive", RHS = Regular)

	DF	SS	% SS	MS	Pr > F	SS	% SS	MS	Pr > F
200901									
AREA	28	8550	0.56%	305.4	0.0311	8172	0.55%	291.9	0.0358
ITEM	210	352131	23.24%	1676.8	<.0001	345557	23.46%	1645.5	<.0001
ERROR	5880	1153900	76.16%	196.2		1119232	75.99%	190.3	
TOTAL	6118	1515086	[100.0%]	SE=	0.1791	1472961		SE=	0.1764
				SEcpi=	0.1068				
200902									
AREA	28	6950	0.67%	248.2	0.0013	8076	0.80%	288.4	<.0001
ITEM	210	314675	30.19%	1498.5	<.0001	297518	29.33%	1416.8	<.0001
ERROR	5880	727671	69.82%	123.8		708788	69.87%	120.5	
TOTAL	6118	1042261	[99.33%]	SE=	0.1422	1014382		SE=	0.1404
				SEcpi=	0.1046				
200903									
AREA	28	6449	0.59%	230.3	0.0016	4762	0.45%	170.1	0.0465
ITEM	210	400684	36.79%	1908.0	<.0001	378377	35.89%	1801.8	<.0001
ERROR	5880	684981	62.89%	116.5		671016	63.65%	114.1	
TOTAL	6118	1089210	[99.73%]	SE=	0.1380	1054155		SE=	0.1366
				SEcpi=	0.1103				
200904									
AREA	28	8609	0.78%	307.5	<.0001	4849	0.45%	173.2	0.0554
ITEM	210	389369	35.13%	1854.1	<.0001	378871	35.06%	1804.1	<.0001
ERROR	5880	711957	64.24%	121.1		697017	64.49%	118.5	
TOTAL	6118	1108323	[99.85%]	SE=	0.1407	1080737		SE=	0.1392
				SEcpi=	0.1041				
200905									
AREA	28	7337	0.64%	262.0	0.0003	6760	0.61%	241.4	0.0008
ITEM	210	433513	37.82%	2064.3	<.0001	420066	37.62%	2000.3	<.0001
ERROR	5880	707378	61.71%	120.3		689660	61.77%	117.3	
TOTAL	6118	1146371	[99.84%]	SE=	0.1402	1116486		SE=	0.1385
				SEcpi=	0.1022				
200906									
AREA	28	6113	0.56%	218.3	0.0028	4706	0.44%	168.1	0.0443
ITEM	210	449713	40.85%	2141.5	<.0001	410938	38.22%	1956.8	<.0001
ERROR	5880	674000	61.22%	114.6		659549	61.34%	112.2	
TOTAL	6118	1100885	[97.44%]	SE=	0.1369	1075193		SE=	0.1354
				SEcpi=	0.0932				

	DF	SS	% SS	MS	Pr > F	SS	% SS	MS	Pr > F
200907									
AREA	28	6191	0.38%	221.1	0.3399	6680	0.41%	238.6	0.2097
ITEM	210	464348	28.17%	2211.2	<.0001	446967	27.64%	2128.4	<.0001
ERROR	5880	1193146	72.39%	202.9		1163581	71.95%	197.9	
TOTAL	6118	1648302	[99.08%]	SE=	0.1821	1617228		SE=	0.1798
				SEcpi=	0.1171				
200908									
AREA	28	5626	0.36%	200.9	0.4729	7621	0.49%	272.2	0.2097
ITEM	210	401047	25.47%	1909.7	<.0001	397458	25.40%	1892.7	<.0001
ERROR	5880	1187991	75.44%	202.0		1159947	74.12%	197.3	
TOTAL	6118	1574770	[98.75%]	SE=	0.1817	1565026		SE=	0.1796
				SEcpi=	0.1022				
200909									
AREA	28	6548	0.59%	233.9	0.0044	6220	0.56%	222.1	0.0063
ITEM	210	350534	31.72%	1669.2	<.0001	368339	33.40%	1754.0	<.0001
ERROR	5880	746430	67.55%	126.9		728126	66.03%	123.8	
TOTAL	6118	1105024	[100.14%]	SE=	0.1440	1102685		SE=	0.1423
				SEcpi=	0.0970				
200910									
AREA	28	3851	0.40%	137.5	0.2145	6176	0.64%	220.6	0.0016
ITEM	210	295500	30.73%	1407.1	<.0001	298946	31.10%	1423.6	<.0001
ERROR	5880	673586	70.05%	114.6		656205	68.26%	111.6	
TOTAL	6118	961640	[98.80%]	SE=	0.1368	961328		SE=	0.1351
				SEcpi=	0.0915				
200911									
AREA	28	4907	0.60%	175.2	0.0063	4971	0.62%	177.5	0.0042
ITEM	210	249101	30.43%	1186.2	<.0001	232602	29.00%	1107.6	<.0001
ERROR	5880	574655	70.20%	97.7		564481	70.38%	96.0	
TOTAL	6118	818603	[98.70%]	SE=	0.1264	802053		SE=	0.1253
				SEcpi=	0.1155				
200912									
AREA	28	3653	0.40%	130.5	0.1685	4769	0.53%	170.3	0.0155
ITEM	210	314418	34.11%	1497.2	<.0001	293139	32.56%	1395.9	<.0001
ERROR	5880	612310	66.42%	104.1		602329	66.91%	102.4	
TOTAL	6118	921844	[99.10%]	SE=	0.1305	900237		SE=	0.1294
				SEcpi=	0.1156				

We can now finally turn to the "adaptive" ANOVA results themselves. In the "**% SS**" column the percentage of sums of squares for AREA in the model never rises above 1%, in fact, averages just above 0.5%. The percentage of sums of squares for ITEM, on the other hand, averages above 30%. But we have to turn to the F tests to find meaningful statistical significance in these numbers. Due to the adaptive nature of the ANOVAs we cannot claim that any of the F test results are precisely correct, but if we can believe in the model fit results in general, then we can accept the F test results as good approximations. To that end, we see that all of the ITEM *p*-values from their F tests are quite simply *zero*. ITEM is always a significant effect in the model. But it is the AREA effect that we are most concerned with in this study. There, 4 of the 12 AREA *p*-values are clearly *not* significant. The other 8 AREA *p*-values *are* significant, at an $\alpha = .025$ level, but with only one *p*-value out of the twelve defined as a zero. This is a mixed result, and does not easily call for the elimination of AREA from the model as a significant main effect. Its "**% SS**" is indeed quite small, but still not so small as to be not significant in two out of every three models examined. Therefore, AREA cannot be ruled out as a significant contributor to the total variance in the CPI model.

An additional two years of ANOVA results were run, and the overall significance results were similar: in 2008, 5 out 12 months showed AREA to be a non-significant effect, in 2007, 3 out of the 12 months showed AREA to be a non-significant effect in the model. Thus, the 2/3 significant, 1/3 non-significant pattern persists through 24 additional months of results

AREA Percentage Variance in the Non-Self-Representing Index-Areas

Turning to the non-self-representing sector of the CPI, we could not adopt an ANOVA methodology in the same way as we have done with A000 (the A-Sized Cities). The seven non-self-representing Index-PSUs *could* be treated as seven more singular AREAs in a larger ANOVA model, but more properly, the actual AREAs (PSUs) in these seven sectors are the multiple smaller PSUs that are contained within each of these Index-PSUs. We do not calculate price relatives for these smaller PSUs for any ITEM or set of ITEMs combination. What we can do, however, is determine components of variance for the random terms in the model.

The generic linear model $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ can consist of fixed effects ($\boldsymbol{X}$) along with any random effects contained within the error term ε. With the A000-SA0 model we had fixed effects for both AREA and ITEM, along with one ERROR term. If we now treat the model as containing only the random effects within the error term ε, we can obtain a different but comparable set of variance *components* within the model. In the non-self-representing Index-Areas we are able to identify as random effects within the model AREA and ITEM (plus OUTLET now and, of course, any remaining ERROR term). We obtain these variance components using a Restricted Maximum Likelihood (REML) methodology. In an earlier memorandum by the author ("Estimation of Variance Components of the U.S. CPI Sample Design" in 1999, and updated in 2008), the theoretics and implementation for using REMLs to produce sets of variance components for the CPI are laid out and explained. For the purposes here, we will simply draw on the variance component results drawn from CPI micro-data from mid-2005 through mid-2008. These variance component results were then averaged across the 36 months of model results for AREA, ITEM, OUTLET and ERROR. The data were at the micro, or unit, level and used a 6-month price relative as the random variable ($\boldsymbol{Y}$) in the model.

The components of variances were then calculated for each Index-Area by Item-Group. There were the 7 Index-Areas, each with at least two or more small PSUs within each, plus 13 Item Groups, each with at least two or more Item-Strata within each Item-Group. One or more outlets were contained then within each Index-Area—Item-Group combination. Thus were we able to generate variance components for AREA, ITEM, OUTLET and ERROR in these 7 non-self-representing Index-Areas by the 13 Item Groups. The composite summary results expressed in percentage of total variance terms for AREA are as follows:

- The MEDIAN percentage of total variance for AREA across the 91 (7 x 13) Area-Item combinations across the 36 months of results was **1.2%**

- The MEAN percentage of total variance for AREA across the 91 (7 x 13) Area-Item combinations across the 36 months of results was **2.0%**

- The MEAN percentage of total variance for AREA across the 91 (7 x 13) Area-Item combinations across the 36 months of results eliminating one egregious outlier was **1.7%**

The comparable summary statistics for ITEM in the Component of Variance models were: MEDIAN percentage of total variance for ITEM = **10.0%**, with MEAN = **14.3%**. Any summary statistics for OUTLET and/or ERROR are not relevant to this study. They are the **82%-86%** leftover in the total variance.

We can compare these variance component summary results with the —%SS" for AREA in the earlier fixed effects models. There the average percentage of total variance for AREA in the self-representing areas was **0.544%**, as compared to the **1.2-2.0%** levels we find in the random effects model in the non-self-representing areas of the CPI. Due to the summary nature of the VC statistics we are presenting here, we cannot provide accurate *p*-values for them. But since the variance contribution of AREA in the non-self-representing AREAs seems to be running at more than twice the percentage level as in the self-representing AREAs, we can only infer that, while AREA is the smallest of all the variance components (in either model, using the term loosely), it is most probably a statistically *significant* term in the model, thus giving added weight to the proposition that AREA *is* an important statistical consideration in any CPI structure that is estimating price change.

Conclusion

In the main effects models, we have found 1/3 of the ANOVA models exhibiting *no* significance whatsoever, yet with the remaining 2/3 of them showing AREA (for the A-Sized Cities sector of the CPI) to be significant, at an $\alpha = .05$ level. Moreover, the component of variance for AREA in the rest of the CPI when treated as a random effect seems to contribute more than twice the percentage of total variance as in the main effects model. However, these percentages are quite small – roughly **1.5%** in the smaller PSUs and roughly **0.5%** in the larger PSUs. AREA comes close to being statistically non-significant, but holds strong enough to claim its place in the overall model.