The Stochastic Approach to Index Numbers: a Misconception

Peter von der Lippe 1/28/2015

Introduction

In what follows[1] I discuss some problematic aspects of the Stochastic Approach in Index Theory (the New Stochastic Approach NSA[2] in particular):

- the NSA unjustly pretends to promote *a better understanding of price index (PI) formulas* by showing that they may be viewed as regression coefficients,
- in both approaches, OSA and NSA prices are assumed to be collected in a random sample; which is not only conspicuously at odds with official price statistics, but triggers much confusion about distributional assumptions allegedly needed for NSA regression models,
- the main advantage of the NSA is often seen in providing an *interval estimate* of a PI in addition to a point estimate, but we doubt that this is of any practical use, and that
- the ostensible usefulness of the NSA for making a choice among index formulas amounts to confounding goodness of fit (related to *data*) with adequacy of a PI formula (which is a *conceptual* and measurement problem, not a problem of fitting the data).

1. **Old (ONA) and new stochastic approach (NSA)**: Jevons and Edgeworth are commonly viewed as "founders" of the OSA in the 19th century. They both advocated with great vigour the *geometric* mean of price relatives that is Jevons' index formula P^J by contrast to Laspeyres who vehemently defended the *arithmetic* mean of price relatives, known as Carli's formula P^C.[3] The concern of the NSA on the other hand is regression models to "explain" a PI as regression coefficient. In some cases, however, such models boil down to a simple restating of PI formulas already known. For example it can easily be seen that (the least squares $\{\Sigma\varepsilon^2 \rightarrow \min\}$ or maximum likelihood) estimator $\hat{\theta}$ for θ in

(1) $\frac{p_{it}}{p_{i0}} = \theta_t + \varepsilon_{it}$ (commodities i = 1, … , n) represents the unweighted arithmetic mean of the price relatives y_{it}, also known as price index formula of Carli

(2) $\hat{\theta}_t = \bar{y} = \frac{1}{n}\sum_i \frac{p_{it}}{p_{i0}} = P_{0t}^C$. For those acquainted with P_{0t}^C the only novelty[4] is that P_{0t}^C can be viewed as $\hat{\theta}$. From this, however, they will learn nothing about the properties of the P^C index (that for example P^C fails the time reversal test). Note also that (1) – as opposed to (3) – not even represents a *regression* model with explicit regressors ("explanatory" or "independent" variables), and that the price relatives p_{it}/p_{i0} in (1) are also viewed as random variable.

Interestingly the modification

[1] This is a shortened version of P. von der Lippe, The Stochastic Approach to Index Numbers: Needless and Useless, MPRA paper number 60839, Dec. 2014..

[2] The CPI Manual of the International Labour Office (ILO) in cooperation with IMF, OECD, UNEC, Eurostat and The World Bank (Geneva 2004) made a distinction between the early "unweighted" and the new "weighted" stochastic approach. We prefer to speak of an old (OSA) and a new (NSA) stochastic approach respectively.

[3] Even Laspeyres made much more use of this formula than of his own formula P^L, which (so it seems) was for him only a more or less unimportant alternative to P^C, at that time the most popular index formula. What is now seen as a major "flaw" of P^C, viz. the absence of "weights" to reflect different "relative importance" of goods was not yet an issue in Laspeyres' days. For more details and the Jevons-Laspeyres controversy see P. von der Lippe, Recurrent Price Index Problems and Some Early German Papers on Index Numbers, Jahrbücher für Nationalökonomie und Statistik (Journal of Economics and Statistics), vol. 233/3 (2013), pp. 336-366.

[4] It is questionable whether this really is a novelty as such an interpretation (the general inflation rate as an average of price relatives) was already quite popular in the Old Stochastic Approach (OSA) of the 19th century.

(1a) $\ln\left(\frac{p_{it}}{p_{i0}}\right) = \theta_t^* + \varepsilon_{it}$ yields Jevons' index $\exp(\hat{\theta}^*) = \prod_i (p_{it}/p_{i0})^{1/n} = P_{0t}^J$. It is telling that already such minor variations concerning the variables of the model will yield quite different index formulas. The same is true for slightly modified assumptions about the error term.

2. **Better understanding of formulas owing to regression models?** The so called NSA – advanced in particular by Clements and Izan (1987)[5], or Selvanathan and Rao (1994)[6] – explicitly boasts itself of enhancing the understanding of index formulas like $P_{0t}^L = \Sigma p_{it}q_{i0}/\Sigma p_{i0}q_{i0}$ (Laspeyres) or $P_{0t}^P\ \Sigma p_{it}q_{it}/\Sigma p_{i0}q_{it}$ (Paasche) by showing that such functions, may be viewed as regression coefficients β_L and β_P in a simple *homogeneous* (or *restricted*, i.e. with the restriction of no intercept, $\alpha = 0$) linear regression equation $y_{it} = \beta_L x_{it} + u_{it}$

(3) $$\frac{p_{it}}{p_{i0}}\sqrt{w_{i0}} = \beta_L \sqrt{w_{i0}} + u_{it}, \qquad u_{it} = \varepsilon_{it}\sqrt{w_{i0}} \quad (i = 1, 2,\ldots, n, \text{ S+P, p. 52})$$

where $w_{i0} = p_{i0}q_{i0}/\Sigma p_{i0}q_{i0}$ are (expenditure) weights and the disturbances ε_{it} are assumed to comply with the standard assumptions $E(\varepsilon_{it}) = 0$, $\mathrm{var}(\varepsilon_{it}) = \sigma^2$, and $\mathrm{cov}(\varepsilon_{it}\varepsilon_{jt}) = 0$ for all $i \neq j$ is assumed (which is certainly not realistic an assumption). As is well known P_{0t}^L is given by

(3a) $$\hat{\beta}_L = \frac{\sum x_{it}y_{it}}{\sum x_{it}^2} = \sum y_{it}x_{it} = \sum \frac{p_{it}}{p_{i0}} w_{i0} = P_{0t}^L,$$

because $\sum x_{it}^2 = \sum\left(\sqrt{w_{i0}}\right)^2 = 1$, while Σx, Σy and Σy^2 are meaningless expressions. The queer factor $\sqrt{w_{i0}}$ makes sure that $P_{0t}^L = \hat{\beta}_L$ is an unbiased estimator of β_L. By contrast, $\hat{\beta}_L^*$ in

(4) $$\frac{p_{it}}{p_{i0}}\sqrt{w_{i0}} = \alpha_L + \beta_L^*\sqrt{w_{i0}} + u_{it}^*$$ is given by (4a) $$\hat{\beta}_L^* = \frac{\frac{1}{n}P_{0t}^L - \bar{x}\cdot\bar{y}}{\frac{1}{n} - \bar{x}^2}.$$

Deriving $\hat{\beta}_L$ in (3a) we see that[7] (deleting subscripts $_{it}$)

(3b) $$\sum \hat{u}^2 = \sum y^2 - \left(\hat{\beta}_L\right)^2 = \sum y^2 - \left(P_{0t}^L\right)^2$$ in contrast to the unrestricted model

(4b) $$\sum \hat{u}^{*2} = \sum y^2 - \left(\hat{\beta}_L\right)^2 - n\hat{\alpha}^2 + (\hat{\beta}_L^* - \hat{\beta}_L)^2 = \sum \hat{u}^2 - n\hat{\alpha}^2 + (\hat{\beta}_L^* - \hat{\beta}_L)^2 \neq \sum \hat{u}^2 .$$[8]

(4b) is a variance decomposition, (3b) is not.[9] Obviously the restricted regression is not a very useful model to "explain" a price index, to compile a confidence interval and to study the goodness of fit (using R^2). Moreover we should ask: Does the fact that a formula, like P^L (or also P^P "explained" by (3) using $(w^*_{i0})^{1/2} = (p_{i0}q_{it}/\Sigma p_{i0}q_{it})^{1/2}$ instead of $(w_{i0})^{1/2}$) may also be interpreted as regression coefficient enable us to develop a better *understanding* of the use and properties of formulas like P^L and P^P as regards for example their axiomatic performance or the uses made of them in official statistics for the purpose of price level measurement (P^L) and deflation (in the case of P^P)? I think, it is most unlikely to know better what P^L really

[5] Clements and Izan, The Measurement of Inflation: A Stochastic Approach, Journal of Business and Economic Statistics, vol. 5, nr. 3 (1987), pp. 339 - 350 (henceforth *quoted as C+I*).

[6] Selvanathan E.A. and D.S. Prasada Rao, Index Numbers, A Stochastic Approach, Ann Arbour (The University of Michigan Press 1994 For the references and more details about much of what is said above I also refer to my book "Index Theory and Price Statistics", Frankfurt/Main 2007, pp. 78 - 90 (in what follows *quoted as S+P*).

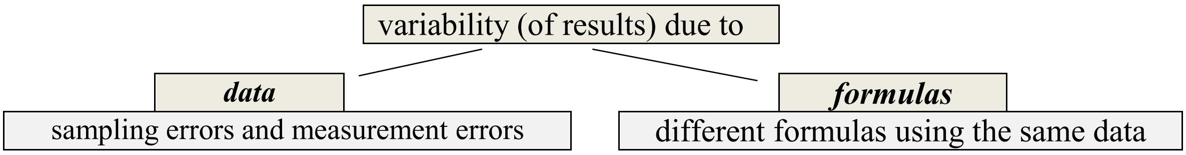
[7] As Σy^2 is not a meaningful expression it would be difficult to give an interpretation to $\Sigma u^2 = \Sigma y^2 - \beta^2$. It is useful now – using the hat notation – to make a distinction between the "true" u and the sample estimate of u.

[8] Also as a rule, Σu (unlike Σu^*) will not vanish (except when $p_{it}/p_{i0} = 1$ for all i, where also $P^L = 1$).

[9] The *variance* of y (*not* Σy^2) is equal to the sum of an explained and a residual component. It is well known that this and R^2, the coefficient of determination does not apply to homogeneous regression.

means when we refer to a regressor $\sqrt{w_{i0}}$. And it is also possibly no coincidence that there seems to be no sensible new index formula owing its discovery to the regression-model-technique of the NSA.

3. **Measures of the adequacy of index formulas** It is said that the NSA provides a measure of the relative adequacy (correctness or appropriateness) of an index formula (which performs better/worse). In our view it is doubtful that a confidence interval (CI), a coefficient of determination $R^2 \approx 1$, or a significance test of a parameter β will be able to serve as such a measure. The fundamental objection is that here the following distinction should be made[10]


variability (of results) due to

data

sampling errors and measurement errors

formulas

different formulas using the same data

Only the data-type of variability is involved when a CI or R^2 is computed. However, to assess the suitability of an index formula requires studying the second type (i.e. the formulas type) of variability, so that we have to compute *different formulas using the same data*. But not only

- should ***fitting data*** (random *variability of observations*)[11] be kept distinct from adequacy, ***correctness*** or so ***of index formulas***, we also should see that
- there is ***nothing that can be inferred from different results*** *when* different index formulas are *applied to the same data*. Such calculations will not give useful information (as regards "the reliability of the index construction" [S+P]) because for *unspecified data (almost*[12]*) any* results may be acceptable. To see this consider the following example

t	1	2	3	4	5
p_{it}/p_{i0}	1,1	1,15	1,18	1,2	1,22

We get $P_{0t}^{C} = 1.17$ and $P_{0t}^{J} = 1.16924$, but we think it leads to nowhere to simply compare such numerical results of index computations for the same data. What consequences (or conclusions) can and should be drawn for example from the fact that $P_{0t}^{C} > P_{0t}^{J}$ = ? This is by the way *necessarily* the case, because – as is well known – for geometric $\bar{x}_G$ and arithmetic means $\bar{x}$ always $\bar{x}_G \leq \bar{x}$ holds) Nothing can be deduced from numerical results of index formulas as such, and it is just for that reason that only *if-then-statements* are used in the axiomatic index theory: *if* no price changes ($p_{it}/p_{i0} = 1 \; \forall i$) *then* the index should display $P_{0t} = 1$ as well.

Experience also shows that very different formulas may yield surprisingly similar numerical results (especially in the case of an only moderately increasing price level).[13]

4. **Thinking in terms of samples and random variables**: The stochastic approach (unlike the axiomatic and the "(micro) economic" approach - both deterministic) requires

a) a specific way of collecting empirical price quotations (observations need to be generated by *random* samples [otherwise there is no point in compiling confidence intervals]), and

[10] The NSA claims to provide *uno actu* a confidence interval CI ("sampling errors … is the concern of the stochastic approach") of an index and a measure of the appropriateness of an index formula (the "stochastic approach therefore concentrates on the search of formulas") apparently by means of R^2.

[11] It could for example turn out that P^L proves the most appropriate formula for a consumer price index in Poland given the time series of prices 2000 – 2010 in Poland while a regression analysis (prices of 2005 – 2014) for the UK shows that P^P would be the best choice. To decide this way on index formulas would of course be ridiculous.

[12] Provided they do not violate some really fundamental axioms such as the mean value property.

[13] I already became aware of such possibilities by some numerical examples in my book "Index Theory…", p. 86. We also should ask: how can we decide over index formulas in the case of overlapping confidence intervals?

b) a number of notoriously misunderstood distributional assumptions.[14]

The problem posed with a) already becomes apparent once we simply ask: From which population and which sampling frame the sample in price statistics is drawn (if there where such a *random* sample at all)? What is the "population" and what its size N? Is it a set of outlets, commodities, acts of purchases/sales, or prices?[15] Is the sample size n the number of commodities and "their" prices (over which summation takes place in the traditional index formulas), or is it the usually much greater number of price quotations collected (in the well known index formulas it is usually assumed that each commodity is represented by only one price).[16]

Moreover as a CPI is compiled on a monthly basis, so the "stochastic" approach should in principle require also *a new random sample drawn every month.* Assume among vegetables the sampling frame contains apples. So there should be months with apples, and those (as it is a *random* selection) without apples, in a similar vein those with shoes and without shoes.

Furthermore, the NSA clearly assumes "disturbance" terms such as ε, u or u* to be *random variables* (and normally distributed for example). Does this also imply the dependent variable y (or even a regressor x) to be *distributed according to a specified probability distribution*? If so, the approach may easily contradict empirical facts. In regression analysis it is uncommon to make assumptions concerning the distribution of empirical data (x variables and y variable) and to refrain from running a regression $y_{it} = \alpha + \beta x_{it} + u_{it}$ (where for example $y_{it} = p_{it}/p_{i0}$) only because y is not normally distributed (as is u).

5. **Sampling distribution and the usefulness of confidence intervals**: Returning to model (1) for Carli's index P^C it is clear that the model entails the (asymptotically normal) sampling distribution $N(\theta, \sigma_y/\sqrt{n})$ for $\bar{y} = \hat{\theta} = \hat{P}^C$, with n as sample size,[17] so that a confidence interval (CI) is easily been calculated.[18] The point now only is:

a) what is the benefit of being able to dispose of a confidence interval (CI), and
b) do we need NSA in order to be able to compile a CI?

To begin with a) we see: From a practical point of view (again from the perspective of official price statistics) we may well question

- Will the ordinary user of CPI statistics see any benefit in a CI of the CPI in addition to the familiar point estimate? Likewise: will monetary policy find more orientation in a *range of probable inflation rates rather than in a single inflation rate*?
- What is the percentage change of P^C from t to t+1 when we have at both times, t and t+1 an *interval* of probable results for $\hat{P}^C_{0,t}$ and $\hat{P}^C_{0,t+1}$ (rather than only one figure, $\hat{P}^C_{0,t}$ and $\hat{P}^C_{0,t+1}$ respectively, in which case the change simply is $\hat{P}^C_{0,t+1}/\hat{P}^C_{0t}$)? Furthermore
- Can we actually increase the sample size (whatever that means) n in order to make the interval smaller and/or to obtain a better estimate of $\hat{\sigma}_y$ in the case of a PI?

[14] Do they apply to the disturbance term or also to the y variable (or even to the x variable(s))?

[15] In other words, sampling (selection) may refer to other units than the subsequent analysis, and also selection can be performed by ways of random selection, or using other methods (non-random selections).

[16] We refer here to the long neglected "low level aggregation" problem (which only recently gained attention), that is the simple (without weights) averaging of *a number of* price quotations for the same good. In practice it is only after this first, or "low" level, that formulas for (weighted) price indices come into play.

[17] See above §4 for the problem on which grounds, if any, we can claim to have a "random sample" and what in particular is the sample size n in this case.

[18] The problem is to estimate σ_y, but once we have sampled values $y_1, y_2, \ldots, y_n$.we also have σ_y.

- What could be a reasonable hypothesis concerning $\theta = P^C$ for example and on which ground can we distinguish θ_0 (hypothesis H_0) and θ_1 (according to H_1) and specify acceptable/desirable levels of the corresponding error risks, α and β respectively. Finally
- In all other parts of official statistics (for example in Population Surveys or National Accounts) it is totally uncommon to refer to random errors.

Now to b): To derive a confidence interval (CI) requires a *sampling* distribution of an *estimator* like $\hat{P}$. It is requisite to know the function $\hat{P}$, which is a linear function in the case of $\hat{P} = P_{0t}^{L} = \sum_i w_{i0}(p_{it}/p_{i0})$. There is no need to know in the context of which regression model we have a regression coefficient β representing P_{0t}^{L}. So once we have a sample (that is the p_{it}/p_{i0} and their variance) there is no need to "explain" P^L in terms of a regression model in order to compile a CI. It is not the regression model which allows (or only facilitates) the derivation of the sampling distribution and thus to estimate a CI. Once we really have a random sample of price relatives we can do without NSA models in order to get an interval estimate.

6. **Price index as regressor or as dependent variable:** Our final argument is that the NSA makes use of regression in a highly unorthodox and not at all sensible way. The "usual" or predominant use made of regression analysis is:

- to identify factors influencing prices in t= 0,1… (or price indices P_{01}, P_{02},…) as the "dependent" variable y_{it}, (I = 1,…,n commodities) that is to gain insights in *data* or to "explain" y_{it}; not to find models where regression coefficients have a specific interpretation;
- to decide on inclusion/exclusion of regressors (x-variables) on the basis of economic theory without any prepossessions as to the "results" we will get for the $\hat{\beta}_k$ coefficients

By contrast in the NSA an entirely different use of the method is made: a model now is devised for the sole purpose of getting a specified function for a *regression coefficient* which may represent a reasonable PI-formula rather than of explaining the "dependent" variable y.

References

Clements, Kenneth W. and H. Y. Izan, The Measurement of Inflation: A Stochastic Approach, Journal of Business and Economic Statistics, vol. 5, nr. 3 (1987), pp. 339 – 350.

International Labour Office (ILO) in cooperation with IMF, OECD, UNEC, Eurostat and The World Bank, CPI Manual, Geneva 2004.

Selvanathan E.A. and D.S. Prasada Rao, Index Numbers, A Stochastic Approach, Ann Arbour 1994.

von der Lippe, Peter, "Index Theory and Price Statistics", Frankfurt/Main 2007.