BLS WORKING PAPERS

U.S. DEPARTMENT OF LABOR
Bureau of Labor Statistics

OFFICE OF PRICES AND
LIVING CONDITIONS

Effects of Mismeasuring Base Period Prices When Estimating the Laspeyres Index: Some Idealized Cases

Timothy Erickson

Working Paper 284
May 1996

The views expressed are those of the author and do not reflect the policies of the U.S. Bureau of Labor Satistics or the views of other BLS staff members.

Effects of Mismeasuring Base Period Prices When Estimating the Laspeyres Index: Some Idealized Cases

Timothy Erickson
Bureau of Labor Statistics

Abstract

The modified Laspeyres price indexes computed by the Bureau of Labor Statistics require knowledge of the base period quantities of each item. Currently, however, only base period expenditure information is available. The Bureau divides these expenditures by estimates of the corresponding base period prices, thereby obtaining estimates of the required quantities. The use of estimated, rather than actual, base period prices is a potential source of error in the resulting index. The magnitude and direction of this error depend on the method used to estimate the base period prices. This note analyzes the impact of alternative methods. For tractability, the analysis assumes a simpler sampling environment than the Bureau actually confronts.

Effects of Mismeasuring Base Period Prices When Estimating the Laspeyres Index: Some Idealized Cases

Timothy Erickson
Bureau of Labor Statistics

October 31, 1995
revised, December 12, 1995

1. Introduction

The modified Laspeyres price indexes computed by the Bureau of Labor Statistics (BLS) require knowledge of the base period quantities of each item. Currently, however, only base period expenditure information is available. BLS divides these expenditures by estimates of the corresponding base period prices, thereby obtaining estimates of the required quantities. The use of estimated, rather than actual, base period prices is a potential source of error in the resulting index. The magnitude and direction of this error depend on the method used to estimate the base period prices. Reinsdorf (1994) has shown that the method used by BLS for many years creates a particular type of upward error known as "formula bias." In this note I analyze the impact of some alternative methods.

For tractability, I assume a simpler sampling environment than the Bureau actually confronts. In particular, I consider the following scenario: suppose we have n independently and identically distributed observations on (B,L,C), and our goal is to estimate $E(C/B)/E(L/B)$. A consistent estimator is the corresponding ratio of the sample means, $\overline{(C/B)}/\overline{(L/B)}$, where $\overline{(C/B)} \equiv n^{-1}\sum_{i=1}^{n} C_i/B_i$ and $\overline{(L/B)} \equiv n^{-1}\sum_{i=1}^{n} L_i/B_i$. To relate this to BLS activities, think of B_i, L_i, and C_i as, respectively, the base, link, and comparison period prices for item i. Under the *i.i.d.* sampling assumption the target quantity $E(C/B)/E(L/B)$ then equals the population Laspeyres index, while the estimator $\overline{(C/B)}/\overline{(L/B)}$ equals the sample Laspeyres index.

3. Process 2

I next derive the bias for Process 2. I first determine the bias in estimating the *numerator* of the target quantity $E(C/B)/E(L/B)$. To do this, I expand C/B around the point $E(C)/E(B)$:

$$\frac{C}{B} \equiv \frac{E(C)}{E(B)} + \frac{\partial}{\partial C}\left(\frac{C}{B}\bigg|_{C=E(C),B=E(B)}\right)(C-E(C)) + \frac{\partial}{\partial B}\left(\frac{C}{B}\bigg|_{C=E(C),B=E(B)}\right)(B-E(B))$$

$$+ \frac{1}{2}\frac{\partial^2}{\partial C^2}\left(\frac{C}{B}\bigg|_{C=E(C),B=E(B)}\right)(C-E(C))^2 + \frac{1}{2}\frac{\partial^2}{\partial B^2}\left(\frac{C}{B}\bigg|_{C=E(C),B=E(B)}\right)(B-E(B))^2$$

$$+ \frac{\partial^2}{\partial C \partial B}\left(\frac{C}{B}\bigg|_{C=E(C),B=E(B)}\right)(C-E(C))(B-E(B))$$

$$= \frac{E(C)}{E(B)} + \frac{1}{E(B)}(C-E(C)) - \frac{E(C)}{[E(B)]^2}(B-E(B))$$

$$+ \frac{E(C)}{[E(B)]^3}(B-E(B))^2 - \frac{1}{[E(B)]^2}(C-E(C))(B-E(B)).$$

Taking the expected value of both sides gives

$$E\left(\frac{C}{B}\right) \equiv \frac{E(C)}{E(B)} + \frac{E(C)}{[E(B)]^3}E\left[(B-E(B))^2\right] - \frac{1}{[E(B)]^2}E\left[(C-E(C))(B-E(B))\right]$$

$$= \frac{E(C)}{E(B)} + \frac{E(C)}{[E(B)]^3}\operatorname{var}(B) - \frac{1}{[E(B)]^2}\operatorname{cov}(C,B)$$

$$= \frac{E(C)}{E(B)} + \frac{E(C)\operatorname{var}(B) - E(B)\operatorname{cov}(C,B)}{[E(B)]^3}.$$

Now assume instead that we do not observe B, but rather $\hat{B}$. I consider a variety of different measurement processes for generating $\hat{B}$. The point of this note is to determine, for each process, the asymptotic bias incurred by estimating $E(C/B)/E(L/B)$ with $\overline{(C/\hat{B})}\big/\overline{(L/\hat{B})}$, where $\overline{(C/\hat{B})} \equiv n^{-1}\sum_{i=1}^{n} C_i/\hat{B}_i$ and $\overline{(L/\hat{B})} \equiv n^{-1}\sum_{i=1}^{n} L_i/\hat{B}_i$.

The processes I consider are:

Process 1: $\hat{B}_i = \varphi L_i$

Process 2: $B_i = \hat{B}_i + u_i \qquad E(u_i) = E(\hat{B}_i u_i) = 0$

Process 3: $B_i = \hat{B}_i v_i \qquad v_i > 0, \quad v_i$ independent of $\hat{B}_i$

Process 4: $\hat{B}_i = B_i + \varepsilon_i \qquad E(\varepsilon_i) = E(B_i\varepsilon_i) = 0$

Process 1 represents the current BLS procedure of using adjusted link period prices as estimates of base period prices. Process 2 represents the hedonic regression $B_i = w_i\theta + u_i$, with $\hat{B}_i \equiv w_i\theta$ and $E(w_i u_i) = 0$. Process 3 represents the logarithmic hedonic regression $\ln(B_i) = w_i\theta + u_i$, with $\hat{B}_i \equiv e^{w_i\theta}$, $v_i = e^{u_i}$, $E(u_i) = 0$, and u_i independent of w_i. Note that using $w_i \equiv 1$ in either Process 2 or 3 will give a $\hat{B}$ that is the same for all i, so that $\overline{(C/\hat{B})}\big/\overline{(L/\hat{B})}$ equals the sample Dutot index. Process 4 does not correspond to any procedure currently proposed for BLS adoption; however, it is the standard "classical errors-in-variables" measurement process, so I have included it for completeness. For all four processes I will evaluate the bias

$$(1) \qquad BIAS \equiv \operatorname{plim}\left[\frac{\overline{(C/\hat{B})}}{\overline{(L/\hat{B})}}\right] - \frac{E(C/B)}{E(L/B)}.$$

2. Process 1 and Formula Bias

Process 1 generates the infamous "formula bias." To obtain a more useful expression for this bias, substitute $\hat{B}_i = \varphi L_i$ into $\overline{(C/\hat{B})}\big/\overline{(L/\hat{B})}$:

$$\frac{\overline{(C/\hat{B})}}{\overline{(L/\hat{B})}} = \frac{n^{-1}\sum_{i=1}^{n} C_i/\hat{B}_i}{n^{-1}\sum_{i=1}^{n} L_i/\hat{B}_i} = \frac{n^{-1}\sum_{i=1}^{n} C_i/\varphi L_i}{n^{-1}\sum_{i=1}^{n} L_i/\varphi L_i} = n^{-1}\sum_{i=1}^{n} C_i/L_i$$

$$\equiv \overline{(C/L)},$$

so the first term of (1) equals $\text{plim}\left[\overline{(C/L)}\right]$. The assumed i.i.d. sampling implies $\text{plim}\left[\overline{(C/L)}\right] = E(C/L)$, so

$$(2) \qquad BIAS_1 = E(C/L) - \frac{E(C/B)}{E(L/B)} = \frac{E(C/L)E(L/B) - E(C/B)}{E(L/B)},$$

where the subscript "1" indicates that this formula is valid only for Process 1. A more insightful expression can be obtained by noting that

$$\text{cov}\left[\left(\frac{C}{L}\right),\left(\frac{L}{B}\right)\right] = E\left[\left(\frac{C}{L}\right)\left(\frac{L}{B}\right)\right] - E\left(\frac{C}{L}\right)E\left(\frac{L}{B}\right) = E\left(\frac{C}{B}\right) - E\left(\frac{C}{L}\right)E\left(\frac{L}{B}\right).$$

Substituting this into (2) gives the "formula bias" as

$$(3) \qquad BIAS_1 = \frac{-\text{cov}\left[(C/L),(L/B)\right]}{E(L/B)}.$$

My intuition is that this will often be positive, because for given C and B an increase in L decreases C/L while increasing L/B.

Treating the ratio $C/\hat{B}$ in the same fashion gives

$$E\left(\frac{C}{\hat{B}}\right) \cong \frac{E(C)}{E(\hat{B})} + \frac{E(C)\operatorname{var}(\hat{B}) - E(\hat{B})\operatorname{cov}(C,\hat{B})}{\left[E(\hat{B})\right]^3}.$$

The probability limit of $\overline{(C/\hat{B})}$ equals $E(C/\hat{B})$, so the asymptotic bias of $\overline{(C/\hat{B})}$ as an estimator of $E(C/B)$ is

$$\begin{aligned}
E\left(\frac{C}{\hat{B}}\right) - E\left(\frac{C}{B}\right) &\cong \frac{E(C)\operatorname{var}(\hat{B}) - E(\hat{B})\operatorname{cov}(C,\hat{B})}{\left[E(\hat{B})\right]^3} - \frac{E(C)\operatorname{var}(B) - E(B)\operatorname{cov}(C,B)}{\left[E(B)\right]^3} \\
&= \frac{E(C)\operatorname{var}(\hat{B}) - E(\hat{B})\operatorname{cov}(C,\hat{B})}{\left[E(\hat{B})\right]^3} - \frac{E(C)\operatorname{var}(B) - E(\hat{B})\operatorname{cov}(C,B)}{\left[E(\hat{B})\right]^3} \\
&= \frac{E(C)\left[\operatorname{var}(\hat{B}) - \operatorname{var}(B)\right] - E(\hat{B})\left[\operatorname{cov}(C,\hat{B}) - \operatorname{cov}(C,B)\right]}{\left[E(\hat{B})\right]^3},
\end{aligned}$$

where the second line follows because Process 2 implies $E(B) = E(\hat{B})$. Process 2 also implies $\operatorname{var}(B) = \operatorname{var}(\hat{B}) + \operatorname{var}(u)$ and $\operatorname{cov}(C,B) = \operatorname{cov}(C,\hat{B}) + \operatorname{cov}(C,u)$, letting us further write

$$(4) \qquad BIAS_{2,N} \equiv E\left(\frac{C}{\hat{B}}\right) - E\left(\frac{C}{B}\right) = \frac{E(\hat{B})\operatorname{cov}(C,u) - E(C)\operatorname{var}(u)}{\left[E(\hat{B})\right]^3},$$

where the subscript "$2,N$" is to remind us that this is the bias, under Process 2, in estimating the *numerator* of our target quantity $E(C/B)/E(L/B)$. A similar analysis applies to the *denominator*, yielding

(5) $$BIAS_{2,D} \equiv E\left(\frac{L}{\hat{B}}\right) - E\left(\frac{L}{B}\right) = \frac{E(\hat{B})\operatorname{cov}(L,u) - E(L)\operatorname{var}(u)}{\left[E(\hat{B})\right]^3}$$

as the bias of $\overline{(L/\hat{B})}$ in estimating $E(L/B)$. Putting the two sets of results together lets us determine the asymptotic bias of $\overline{(C/\hat{B})}/\overline{(L/\hat{B})}$ as an estimator of $E(C/B)/E(L/B)$. There are a variety of ways of expressing this bias. One way is obtained by substituting

(6) $$\operatorname{plim}\left[\frac{\overline{(C/\hat{B})}}{\overline{(L/\hat{B})}}\right] = \frac{\operatorname{plim}\left[\overline{(C/\hat{B})}\right]}{\operatorname{plim}\left[\overline{(L/\hat{B})}\right]} = \frac{E(C/\hat{B})}{E(L/\hat{B})} = \frac{E(C/B) + BIAS_{2,N}}{E(L/B) + BIAS_{2,D}}$$

for the first right hand side term in (1), and then rearranging the result as

(7) $$BIAS_2 = \left(\frac{\dfrac{BIAS_{2,N}}{E(C/B)} - \dfrac{BIAS_{2,D}}{E(L/B)}}{\left[1 + \dfrac{BIAS_{2,D}}{E(L/B)}\right]}\right)\left(\frac{E(C/B)}{E(L/B)}\right)$$

In general this is nonzero, but unlike "formula bias" has an indeterminate sign. Note that the bias equals zero if

(8) $$\frac{BIAS_{2,N}}{E(C/B)} = \frac{BIAS_{2,D}}{E(L/B)}.$$ (Zero bias condition for Process 2)

Thus, if we think the numerator and denominator biases are about the same proportion of their targets, then the overall bias (7) should be close to zero. This would appear to be a plausible assumption in at least some instances.

Under Process 2 it may be possible to obtain consistent estimates of the target quantity $E(C/B)/E(L/B)$, even when the estimator $\overline{(C/\hat{B})}/\overline{(L/\hat{B})}$ suffers from a nonzero value for the bias (7). If the measurement error equation $B = \hat{B} + u$ is an out-of -sample predictor obtained from hedonic regression, then there will be residuals $\hat{u}_i$ from estimating the latter which can be used to consistently estimate $\text{var}(u)$, $\text{cov}(C,u)$, and $\text{cov}(L,u)$. Estimates of $E(C)$, $E(L)$, and $E(\hat{B})$ can be obtained from the corresponding sample means. Plugging all of these estimates into (4) and (5) then gives consistent estimates of $BIAS_{2,N}$ and $BIAS_{2,D}$. Substituting the latter along with $\overline{(C/\hat{B})}$ and $\overline{(L/\hat{B})}$ into the right hand side of

$$(9) \qquad \frac{E(C/B)}{E(L/B)} = \frac{E(C/\hat{B}) - BIAS_{2,N}}{E(L/\hat{B}) - BIAS_{2,D}}$$

then gives a consistent estimator. CAUTION: The resulting consistent estimator may well have a larger finite sample mean square error (or other measure of accuracy) than does the inconsistent estimator.

Recall the special case of Process 2 where $w_i \equiv 1$ implies the same $\hat{B}$ for every i, resulting in a Dutot index. Here the first term in (1) satisfies

$$(10) \qquad \text{plim}\left[\frac{\overline{(C/\hat{B})}}{\overline{(L/\hat{B})}}\right] = \frac{\text{plim}\left[\overline{(C/\hat{B})}\right]}{\text{plim}\left[\overline{(L/\hat{B})}\right]} = \frac{E(C)/\hat{B}}{E(L)/\hat{B}} = \frac{E(C)}{E(L)},$$

which when substituted into (1) gives the Dutot bias as

$$(11) \qquad BIAS_{Du} = \frac{E(C)}{E(L)} - \frac{E(C/B)}{E(L/B)}.$$

A zero bias condition analogous to (8) would be

$$(12) \qquad \frac{E(C)}{E(C/B)} = \frac{E(L)}{E(L/B)}. \qquad \text{(Zero bias condition for Dutot)}$$

4. Process 3

To evaluate the bias for Process 3 we write the first term in (1) as

$$(13)\quad \frac{E\left(\frac{C_i}{\hat{B}_i}\right)}{E\left(\frac{L_i}{\hat{B}_i}\right)} = \frac{E\left(\frac{C_i}{B_i/v_i}\right)}{E\left(\frac{L_i}{B_i/v_i}\right)} = \frac{E\left(\frac{v_i C_i}{B_i}\right)}{E\left(\frac{v_i L_i}{B_i}\right)};$$

dropping the i subscripts for convenience, the bias formula (1) becomes

$$(14)\quad BIAS_3 = \frac{E\left(\frac{vC}{B}\right)}{E\left(\frac{vL}{B}\right)} - \frac{E\left(\frac{C}{B}\right)}{E\left(\frac{L}{B}\right)}.$$

Like the bias in Process 2 the algebraic sign of this bias appears indeterminate. The special case where $w_i \equiv 1$ results in the Dutot process, in which case (14) will equal (11).

Modifying Process 3 by further assuming that v_i is independent of (C_i, L_i) will eliminate the bias. This follows from

$$(15)\quad \frac{E\left(\frac{C_i}{B_i}\right)}{E\left(\frac{L_i}{B_i}\right)} = \frac{E\left(\frac{C_i}{\hat{B}_i v_i}\right)}{E\left(\frac{L_i}{\hat{B}_i v_i}\right)} = \frac{E\left(\frac{C_i}{\hat{B}_i}\left(\frac{1}{v_i}\right)\right)}{E\left(\frac{L_i}{\hat{B}_i}\left(\frac{1}{v_i}\right)\right)} = \frac{E\left(\frac{C_i}{\hat{B}_i}\right)E\left(\frac{1}{v_i}\right)}{E\left(\frac{L_i}{\hat{B}_i}\right)E\left(\frac{1}{v_i}\right)}$$

$$= \frac{E\left(\frac{C_i}{\hat{B}_i}\right)}{E\left(\frac{L_i}{\hat{B}_i}\right)}.$$

are consistently estimable. Combining these facts with the appropriate analog to (9) yields the following consistently estimable bounds on the target quantity: if $E(L/\hat{B}) \neq U(BIAS_{4,D})$ then

$$\text{(19)} \quad \frac{E(C/\hat{B})}{E(L/\hat{B}) - U(BIAS_{4,D})} \leq \frac{E(C/B)}{E(L/B)} \leq \frac{E(C/\hat{B}) - U(BIAS_{4,N})}{E(L/\hat{B})}.$$

Unfortunately, this additional assumption is not likely to be true. For example, assume the simplest hedonic specification, $\ln(B_i) = \theta + u_i$, in which case $\hat{B}_i = e^{\theta}$ is a constant not depending on i. All of the variation in B_i is then due to v_i, so to say that the latter is independent of (C_i, L_i) is equivalent to saying B_i is independent of (C_i, L_i). More generally, the additional assumption requires w_i explain all of the variation in B_i that is related to (C_i, L_i).

5. Process 4

The analysis of Process 4 is similar to that for Process 2, and leads to

$$(16) \quad BIAS_{4,N} \equiv \frac{E(C)\operatorname{var}(\varepsilon) - E(\hat{B})\operatorname{cov}(C,\varepsilon)}{\left[E(\hat{B})\right]^3}$$

$$(17) \quad BIAS_{4,D} \equiv \frac{E(L)\operatorname{var}(\varepsilon) - \frac{1}{2}E(\hat{B})\operatorname{cov}(L,\varepsilon)}{\left[E(\hat{B})\right]^3}$$

$$(18) \quad BIAS_4 = \left(\frac{\dfrac{BIAS_{4,N}}{E(C/B)} - \dfrac{BIAS_{4,D}}{E(L/B)}}{\left[1 + \dfrac{BIAS_{4,D}}{E(L/B)}\right]} \right) \left(\frac{E(C/B)}{E(L/B)} \right).$$

Consistent estimation is not possible under Process 4. However, with some additional assumptions on the measurement process it is possible to estimate a type of "errors-in-variables" bound. In particular, assume $\operatorname{cov}(C,\varepsilon) = 0$ and $\operatorname{cov}(L,\varepsilon) = 0$, so that $BIAS_{4,N} \equiv E(C)\operatorname{var}(\varepsilon) \big/ \left[E(\hat{B})\right]^3$ and $BIAS_{4,D} \equiv E(L)\operatorname{var}(\varepsilon) \big/ \left[E(\hat{B})\right]^3$. The original assumptions of Process 4 then imply $0 \le \operatorname{var}(\varepsilon) \le \operatorname{var}(\hat{B})$, and thus $0 \le BIAS_{4,N} \le E(x)\operatorname{var}(\hat{B}) \big/ \left[E(\hat{B})\right]^3 \equiv U(BIAS_{4,N})$ and $0 \le BIAS_{4,D} \le E(L)\operatorname{var}(\hat{B}) \big/ \left[E(\hat{B})\right]^3 \equiv U(BIAS_{4,D})$. Note that both $U(BIAS_{4,N})$ and $U(BIAS_{4,D})$

6. Conclusion

This concludes my formal derivations. It may be useful at this point to mention that the sample geometric mean, *when considered as an estimator of* $E(C/B)/E(L/B)$, is also inconsistent. This is because the sample geometric mean *is* a consistent estimator of the population geometric mean, and the latter generally differs from the population Laspeyres $E(C/B)/E(L/B)$. The algebraic sign of this bias is indeterminate.

I have found that $\overline{(C/\hat{B})}/\overline{(L/\hat{B})}$ is an inconsistent estimator of $E(C/B)/E(L/B)$ except under the (probably unrealistic) *modified* Process 3. As noted, the geometric mean is also inconsistent. A consistent estimator of $E(C/B)/E(L/B)$ that is *not* of the form $\overline{(C/\hat{B})}/\overline{(L/\hat{B})}$ may be available under Process 2. However, this estimator depends on a first stage of bias estimation, and the additional noise arising from this first stage may cause the resulting consistent estimator to have larger finite sample mean squared error than the inconsistent estimator $\overline{(C/\hat{B})}/\overline{(L/\hat{B})}$.

It is important to note that only the "formula bias" in expression (3) could be confidently given an algebraic sign based on prior knowledge. In my opinion an estimator having a bias of unknown sign can for many practical purposes be treated as if it were unbiased. This view is justified by Bayesian analyses using zero-mean prior distributions for the bias, which can result in posterior distributions centered at the realized value for $\overline{(C/\hat{B})}/\overline{(L/\hat{B})}$; in such analyses the unknown bias causes only an increase in posterior uncertainty, not a shift in posterior location. (See Leamer, p. 298.) This explains why many (all?) CPI critics seem exclusively concerned with whether alleged biases are "upward" or "downward." They may well be satisfied with an alternative biased estimator, so long as the algebraic sign of the bias is unknown and/or is thought to be close to zero. My results suggest that this is probably the best we can do, so long as the population Laspeyres is the target parameter. The situation is different if instead the population geometric mean is the target parameter. In the latter case the sample geometric mean is a consistent estimator. Yet another possible target parameter is the "true" cost-of-living index derived from some aggregate utility function. Here data limitations again create a situation in which the best we could hope to do with regard to asymptotic bias is obtain an inconsistent estimator having a bias of unknown sign and/or likely to be close to zero.

A final caveat. This paper has addressed concerns about asymptotic bias. By itself this can only be a useful criterion for estimator performance if it is assumed that competing estimators have essentially identical finite sample biases and (more importantly) variances. In particular, finite sample "closeness" criteria such as mean absolute error, mean squared error, probabilities of lying within a given distance of the target parameter, etc., are superior guides for determining accurate estimates.

References

Leamer, Edward E. (1978): *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Reinsdorf, Marshall (1994): "Price Dispersion, Seller Substitution, and the U.S. CPI," BLS working paper No. 252.