INTERNATIONAL PRICE COMPARISONS BASED UPON INCOMPLETE DATA*

BY ROBERT SUMMERS

University of Pennsylvania

This paper is directed at the following question: given an incomplete set of price data relating to goods or services in some category of output for each of a number of different countries, what arithmetic should be performed on the prices to get a meaningful representation of the relative category price-levels of the countries? In the course of developing an answer to the question, some broader matters are considered and illuminated. A comparison of category price-levels for different countries is analogous to a commonly-encountered problem in many areas, that of ranking ordinally or cardinally in one dimension a group of "entities"—persons, households, firms, industries, etc.—on the basis of sets of measurements associated with the individual entities. It is this point of view which dominates the following presentation.

I. INTRODUCTION

Traditionally, comparisons of price-levels at different times or in different places are made on the basis of relative costs of market baskets, somehow computed, without explicit reference to stochastic distributions and principles of statistical inference. The economic theory underlying such cost-of-living estimates deals with the implications of negatively-sloped demand curves and emphasizes the general importance and difficulties of applying proper weighting procedures.[1] Essential as this theory is, it should be regarded as a complement to rather than a substitute for a stochastic framework. Economic theory may dictate which "statistic" should be computed from the available data, but the position taken here is that a country's category price-level still should be regarded as an appropriately defined center of a stochastic distribution. However, the proper estimate of that center is not necessarily simply an average of prices of individual category items. The primary message here is an obvious one: the price-level should be estimated using a procedure founded upon statistical inference principles. These remarks are not meant to minimize the importance of "the index number

*This paper was prepared as a part of the research program of the International Comparison Project, a joint activity of the United Nations Statistical Office and a group of researchers at the University of Pennsylvania. The Project's support has come primarily from the Ford Foundation, with significant additional financial help being provided by the International Bank for Reconstruction and Development. The many thoughtful discussions of various statistical points the author had with his former colleague, Edward Prescott, are gratefully acknowledged. Irving B. Kravis and Alan Heston of the ICP provided invaluable help in assessing the operational usefulness of the methods developed. Able computing assistance, not all on display in this version of the research, was provided by Lorenzo Perez.

[1]Everyone's list of the classic works on index number theory would include *The Making of Index Numbers* by Irving Fisher (Houghton, Miflin, 1922) and "Annual Survey of General Economic Theory: The Problem of Index Numbers" by Ragnar Frisch (*Econometrica*, Vol. 4, No. 1, pp. 1–39; January 1936). An extremely comprehensive, up-to-date review of the index number literature, with its own independent contribution to the art, is "Price Indexes and International Price Comparisons" by Richard Ruggles in *Ten Economic Studies in the Tradition of Irving Fisher* (Wiley, 1967).

problem"—though it will be ignored here—but rather to set the stage for likelihood functions and their associated statistical apparatus.

The problem of incomplete price data is central to the discussion below. The frequency with which multilateral comparisons must be made on the basis of incomplete binary data in empirical work generally justifies looking carefully at this sort of problem in any case. The specific impetus for concern now is the need of the International Comparison Project of the Statistical Office of the United Nations and the University of Pennsylvania for a procedure whereby price indexes can be computed for each of a large number of narrow categories for eight or ten countries. Despite reasonable diligence, it has not been possible to price all items within each category for every one of the countries. Discarding all price data on items for which any country prices are missing is obviously inefficient; the procedure developed below is designed to utilize all available data in an economical way.

In Section II the multilateral ranking problem is set down, first for complete price information and then for the situation in which some prices are missing. A possible nonstochastic treatment is suggested there before Section III presents a formal—but simplistic—framework for analysis. By the end of Section III a formidable likelihood function is derived. Happily, it can be shown that the underlying stochastic model has a regression interpretation which allows for Section IV's simpler operating procedure. The so-called Country-Product-Dummy method (CPD) is developed and then applied to a set of hypothetical but not wholly implausible data. After the worked example of Section V, brief concluding remarks are given in Section VI.

II. A Non-Stochastic Formulation

(1) $P =$

	Country (i)			
Item (α)	1	2	...	m
1	p_{11}	p_{12}	...	p_{1m}
2	p_{21}	p_{22}	...	p_{2m}
.	.	.		.
.	.	.		.
.	.	.		.
A	p_{A1}	p_{A2}	...	p_{Am}

Consider the price tableau P given by (1). All national prices have been converted to a standardized currency unit, perhaps but not necessarily by an official exchange rate. The units in which the items are measured are arbitrary (e.g., p_{1i} may be the price of potatoes per peck and p_{2i} may be the price of radishes per bunch, etc.). Let Items 1 through A be from some fairly narrowly defined category of goods and services (like vegetables). Suppose because quantities are simply unknown, a comparison of category price-levels is to be made for the m countries *without* using quantity weights. (The error resulting from neglecting quantity weights is likely not to be serious if either the relative quantities or relative prices of the A items are not too dissimilar within each of

the m countries. Whether one or the other of these conditions is met is obviously an empirical matter; for fairly homogeneous categories, there is a good chance that quantity weights may be safely ignored.) "*The*" index number problem is ignored at the category level; the importance of different items is assumed to be the same. It should be noted though that while the insensitivity of category price-level estimates to various weighting schemes may justify proceeding without quantities here, of course the International Comparison Project will face directly the weighting problem when different category price indexes are combined.

The relative price-levels for the m countries may be defined in a number of possible ways. An essential characteristic of any definition is that its implementing formula should give results that are invariant both under a change in the units of the items and also a change in the country used as a base in the comparisons; in addition, the property of "circularity" (sometimes called "transitivity") for multilateral comparisons is very much to be desired.[2]

Perhaps the best-known price index formula (but one which involves weights), the Laspeyres index, is given in (2).

$$(2) \qquad I_{ij} = \frac{\sum_{\alpha=1}^{A} p_{\alpha i} q_{\alpha j}}{\sum_{\alpha=1}^{A} p_{\alpha j} q_{\alpha j}} = \sum_{\alpha=1}^{A} w_\alpha^{(j)} \left(\frac{p_{\alpha i}}{p_{\alpha j}}\right) \quad \text{where } w_\alpha^{(j)} = \frac{p_{\alpha j} q_{\alpha j}}{\Sigma p_{\alpha j} q_{\alpha j}}.$$

If all items are considered equally important in the sense that all the $w_\alpha^{(j)}$'s are equal to $1/A$, then (2) reduces to (2′).

$$(2') \qquad I_{ij}^{(1)} = \sum_{\alpha=1}^{A} \left(\frac{p_{\alpha i}}{p_{\alpha j}}\right) \Big/ A.$$

$I_{ij}^{(1)}$ gives the average relative cost of each item of the same set of goods and/or services in Countries i and j, and it is indeed invariant under a change in units. However, it is not invariant under a change in the country selected as a base, and it does not possess the circularity property. By a simple change, these defects can be remedied: if instead of defining the index as the arithmetic mean of the individual price relatives, $I_{ij}^{(2)}$ is taken to be the *geometric mean*, as in (3), then $I_{ij}^{(2)}$ will meet the unit and base invariance requirements and will possess the circularity property.

$$(3) \qquad I_{ij}^{(2)} = \left[\frac{p_{1i}}{p_{1j}} \cdot \frac{p_{2i}}{p_{2j}} \cdot \cdots \cdot \frac{p_{Ai}}{p_{Aj}}\right]^{1/A}.$$

These remarks about $I_{ij}^{(2)}$ hold when (3) is applied to the full price tableau P. If, however, some prices are missing, clearly all of the item price ratios called for in (3) cannot be calculated. The obvious modification of the formula, given in (3′), would appear to be justified if in some sense the missing prices are

[2]Invariance under a change of units requires that if $I_{ij} = \varphi(P)$ is an estimate of the relative price-levels of Countries i and j, then multiplying all elements in a row of P by the same constant c_α will not lead to a new value of I_{ij}. Invariance under a change of base requires that $I_{ij} = 1/I_{ji}$. (The so-called "time-reversal test" of index number theory refers to this invariance in the context of a time-to-time comparison.) An index is said to possess the circularity property if the binary cardinal rankings of entities i, j, and k are related to each other as follows:

$$I_{ij} = I_{ik}/I_{jk}$$

"typical" ones (i.e., if, loosely speaking, the missing prices are a *random* selection of entries from the tableau).

$$(3') \qquad I_{ij}^{(3)} = \left[\frac{p_{1i}}{p_{1j}} \cdot \frac{p_{2i}}{p_{2j}} \cdots\right) \frac{p_{\bar{\alpha}_1 i}}{p_{\bar{\alpha}_1 j}} \cdots \frac{p_{\bar{\alpha}_n i}}{p_{\bar{\alpha}_n j}} \left(\cdots \frac{p_{mi}}{p_{mj}}\right]^{1/(A-n)}.$$

where $)\ldots($ refers to omitted price ratios and $\bar{\alpha}_1, \ldots, \bar{\alpha}_n$ is the set of n items for which either $p_{\alpha i}$ or $p_{\alpha j}$ (or both) is missing.

(3′) parallels (3) in that $I_{ij}^{(3)}$ is the geometric mean of all *available* price ratios. Though $I_{ij}^{(3)}$ is invariant under changes in units and country base, missing prices in various rows and columns will lead to noncircularity. This means that an indirect comparison of Country i with Country j by means of the ratio $I_{i1}^{(3)}/I_{j1}^{(3)}$, where Country 1 is a base country, will not in general be equal to the direct comparison, $I_{ij}^{(3)}$. Unfortunately, this ambiguity makes it impossible to get a unique cardinal ranking of all m countries. (m–1) numbers, each assigned to a nonbase country, are required for a unique cardinal ranking, but (3′) gives a set of $m(m-1)/2$ numbers which do not line up properly to give a single ranking.

It should be made clear that the point of seeking circularity is not merely to avoid this non-uniqueness. If indeed circularity is present in the real world, then it is possible to supplement the direct information bearing on Country i's price-level relative to Country j's derived from the set of available $(p_{\alpha i}/p_{\alpha j})$'s. Circularity implies that for items where either $p_{\alpha i}$ or $p_{\alpha j}$ is missing, say p_{1i} and p_{2j}, the ratios p_{1j}/p_{1k} and p_{2i}/p_{2k} (i.e., other price ratios involving the same items) provide information bearing upon the missing ratios p_{1i}/p_{1j} and p_{2i}/p_{2j}. Operationally, this means that if circularity can be safely assumed, more precise estimates of relative price-levels can be estimated by taking explicit account of circularity. (That this may not be so will be seen in the numerical example of Section V.)

An intuitive way of patching up the non-circularity would be to estimate the I_{ij}'s using the geometric means of (3′) but somehow to force the estimates to meet the circularity conditions $I_{ij} = I_{ik}/I_{jk}$ for all $i \neq j$. A natural strategy to experiment which might be the employment of a Least Squares procedure in which Q as given in (4) is minimized with respect to the $(m-1)$ price-level ratios (p_i^*/p_1^*).

$$(4) \qquad Q = \left[I_{21}^{(3)} - \frac{p_2^*}{p_1^*}\right]^2 + \ldots + \left[I_{m1}^{(3)} - \frac{p_m^*}{p_1^*}\right]^2 + \left[I_{32}^{(3)} - \frac{p_3^*}{\beta_2^*}\right]^2 \cdots$$
$$+ \left[I_{m2}^{(3)} - \frac{p_m^*}{p_2^*}\right]^2 + \ldots + \left[I_{m,m-1}^{(3)} - \frac{p_m^*}{p_{m-1}^*}\right]^2$$

If p_i^*/p_1^* is denoted θ_{ij} and particularly p_i^*/p_1^* is denoted $\theta_i (\theta_1 = 1)$, then Q can be written more simply as in (4′). Now the minimization should be carried out with respect to $\theta_2, \theta_3, \ldots, \theta_m$.

$$(4') \qquad Q = \sum_{j=1}^{m} \sum_{i=j+1}^{m} \left[I_{ij}^{(3)} - \frac{\theta_i}{\theta_j}\right]^2.$$

Though circularity is achieved here, the minimizing values of the θ_i's are not independent of the numbering of the countries in (4) and (4′). The logic of the

procedure would not rule out having the indexes of the summations interchanged; but if Q were equal to the sum of such terms, starting with $[I_{12}^{(3)} - \theta_1/\theta_2]^2$, the resulting minimizing values of the θ_i's would be different. (Incidentally, the minimization process would require use of a gradient algorithm because the first order conditions give rise to non-linear equations. However, with the high-speed computers available now, this factor alone would not be a major consideration.)

If this loss of the base-invariance property were the only difficulty with this Least Squares method, a simple remedy for (5′) could be found. Minimizing $\bar{Q}$ as given in (5) with respect to $\theta_2, \theta_3, \ldots, \theta_m$ would give price-level relatives with the desired invariances and circularity too.

$$\bar{Q} = \sum_{j=1}^{m} \sum_{i=j+1}^{m} \left[\ln I_{ij}^{(3)} - \ln \left(\frac{\theta_i}{\theta_j} \right) \right]^2. \tag{5}$$

Converting to natural logarithms has the added advantage that the values of the θ_i's which maximize $\bar{Q}$ can be found by solving a set of linear equations (where the unknowns are $\ln \theta_2, \ldots, \ln \theta_m$).[3]

This *ad hoc* development of a procedure for estimating the relative price-levels $\theta_2, \ldots, \theta_m$ using (5) has some intuitive appeal, but it clearly has many arbitrary elements. What "loss function" describing the cost of errors in price-level estimates does $\bar{Q}$ imply? $\bar{Q}$ is quadratic; should it not also have cross-product terms? Since the number of elements, N_{ij}, entering into the geometric mean $I_{ij}^{(3)}$ is not the same for all two-country comparisons, is it reasonable for the coefficients of the individual quadratic terms on the right side of (5) all to be the same? What is the precision of the estimates of the relative price-levels as derived from the Least Squares procedure? Clearly, a more systematic, less handwaving approach is required.

III. A Stochastic Formulation

If the price tableau, P, of (1) is the set of available data about a category, how many items beyond A are there in the category which at least in principle could be priced but in fact have not been? For the purposes of most of what follows, it will be assumed that there is an indefinitely large number of items in each category and that A represents only the number of observations actually observed in a sample which has been generated by the following model.

Pairs of prices in any row of P are assumed to be related to each other as indicated in (6).

$$\frac{p_{\alpha i}}{p_{\alpha j}} = \frac{p_i^*}{p_j^*} \cdot w_\alpha^{ij}. \tag{6}$$

Here (p_i^*/p_j^*) is again the relative price-level in Countries i and j; and w_α^{ij} is a random variable which is lognormally distributed with parameters 0 and σ^2. The assumption that the w_α^{ij}'s are distributed lognormally is a typical one in

[3]This formulation is similar to the "one-dimensional price scale" work of H. Theil in *Economics and Information Theory* (Rand McNally, Chicago; North-Holland, Amsterdam; 1967), pp. 135–150. Theil's procedure (p. 147) calls for minimizing a "sum of squared residuals" where the summation on the right side of (5) covers *all i–j* combinations.

situations where multiplicative relationships are assumed. However, the assumption that σ^2 needs no subscripts or superscripts is a strong one; it asserts that the variance of price ratios is the same for all pairs of countries. Clearly, such an assumption can only be defended on the grounds that the available data are insufficient in quantity to allow a more realistic distinction to be made empirically between country pairs. Fortunately, estimates would be unbiased even if this assumption was violated. More importantly, it is assumed that the $w_\alpha{}^{ij}$ and $w_{\alpha'}{}^{ij}$ are independent for $\alpha \neq \alpha'$. This is the sense in which random sampling is assumed, and this assumption *must* be complied with.

To simplify the stochastic presentation, a set of new variables is defined.

$$R_\alpha{}^{ij} = \ln\left(\frac{p_{\alpha i}}{p_{\alpha j}}\right); R_\alpha{}^{ij} = -R_\alpha{}^{ji} \tag{7}$$

$$\mu_{ij} \equiv \ln\left(\frac{p_i^*}{p_j^*}\right); \mu_{ij} = -\mu_{ji}; \mu_{i1} \equiv \mu_i \tag{7'}$$

$$\epsilon_\alpha{}^{ij} \equiv \ln w_\alpha{}^{ij}; \epsilon_\alpha{}^{ij} = -\epsilon_\alpha{}^{ij}. \tag{7''}$$

In this new notation (6) can be rewritten as (8).

$$R_\alpha{}^{ij} = \mu_{ij} + \epsilon_\alpha{}^{ij} \quad \text{where } f(\epsilon_\alpha{}^{ij}): \quad \text{normal}\,(0, \sigma^2). \tag{8}$$

It follows immediately from (7), (7′) and (7″) that:

$$R_\alpha{}^{ij} + R_\alpha{}^{jk} + R_\alpha{}^{ki} = 0 \tag{9}$$

$$\mu_{ij} + \mu_{jk} + \mu_{ki} = 0. \tag{9'}$$

Therefore, the relationship among the stochastic elements, $\epsilon_\alpha{}^{ij}$, will be as given in (10):

$$\epsilon_\alpha{}^{ij} + \epsilon_\alpha{}^{jk} + \epsilon_\alpha{}^{ki} = 0 \quad \text{or } \epsilon_\alpha{}^{ij} = \epsilon_\alpha{}^{ik} - \epsilon_\alpha{}^{jk}. \tag{10}$$

If Country 1 is thought of as the base country, then particularly

$$\mu_{ij} = \mu_{i1} - \mu_{j1} \equiv \mu_i - \mu_j \tag{11}$$

and

$$\epsilon_\alpha{}^{ij} = \epsilon_\alpha{}^{i1} - \epsilon_\alpha{}^{j1}. \tag{12}$$

(12) appears to be a relationship that precisely parallels (11) and therefore should merely be designated (11′). It deserves separate recognition, however. Notice that if P is complete—i.e., has no "holes"—there will be $A \cdot m(m-1)/2$ different $R_\alpha{}^{ij}$'s. At first sight it might appear that there are that many independent observations bearing on the individual unknown (p_i^*/p_j^*)'s. The relationship in (12) shows, however, that they are not all independent. In fact, $A \cdot (m-1)(m-2)/2$ of the observations are *redundant* in the sense that they are exact linear combinations of $A(m-1)$ basic observations. (This is because any "error term", $\epsilon_\alpha{}^{ij}$, not involving the base country, Country 1, is functionally dependent upon two ϵ's involving the base country. There are $A \cdot m(m-1)/2$ ϵ's in all; there are $A \cdot (m-1)$ ϵ's involving the base country; and there are $A \cdot (m-1)(m-2)/2$ ϵ's not involving the base country, each of which can be expressed as the difference between two base country ϵ's.)

More than that, close examination of (12) reveals that even the $A \cdot (m-1)$ ϵ's involving the base country are not statistically independent. If each side of (12)

is squared and then the expected-value operator is applied to each side, it can be seen that the covariance of $\epsilon_\alpha{}^{i1}$ and $\epsilon_\alpha{}^{j1}$ is $\sigma^2/2$.

For any α with a full row of $p_{\alpha i}$, the functional dependence of all $\epsilon_\alpha{}^{ij}$, $i > j$, and the statistical interdependence of all $\epsilon_\alpha{}^{i1}$ and $\epsilon_\alpha{}^{j1}$ together imply that the joint density function of $\epsilon_\alpha{}^{ij}$ for all $i > j$ can be written simply as the joint density function of the $(m-1)$ $\epsilon_\alpha{}^{i1}$'s. (13) specifies this joint density

(13)
$$f_1(\epsilon_\alpha{}^{21}, \ldots, \epsilon_\alpha{}^{m1}, \epsilon_\alpha{}^{32}, \ldots, \epsilon_\alpha{}^{m2}, \ldots, \epsilon_\alpha{}^{m,m-1}) = f_1(\epsilon_\alpha{}^{21}, \ldots, \epsilon_\alpha{}^{m,1}) : N_{m-1}\{\bar{0}_{m-1}; \sigma^2 V_{m-1}\}$$

where N_k signifies a joint k-variate normal density function; $\bar{0}_k$ is a $(k \times 1)$ vector consisting entirely of zeros; and V_k is a $(k \times k)$ matrix consisting of 1's on the principal diagonal and $\frac{1}{2}$'s everywhere else.

Specifically, $N_{m-1}\{\bar{0}_{m-1}; \sigma^2 V_{m-1}\}$ in this case means:

(14)
$$f_1\left(\epsilon_\alpha{}^{21}, \ldots, \epsilon_\alpha{}^{m,1}\right) = \frac{1}{(2\pi)^{(m-1)/2}} \cdot \frac{1}{\sigma |V_{m-1}|^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\left(\epsilon_\alpha{}^{(1)}\right)' V_{m-1}{}^{-1}\left(\epsilon_\alpha{}^{(1)}\right)\right]$$

where

$$\epsilon_\alpha{}^{(1)} \equiv \begin{pmatrix} \epsilon_\alpha{}^{21} \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_\alpha{}^{m1} \end{pmatrix}.$$

From (8) it follows that for any α:

(15)
$$f_2(R_\alpha{}^{21}, \ldots, R_\alpha{}^{m1}, R_\alpha{}^{32}, \ldots, R_\alpha{}^{m2}, \ldots, R_\alpha{}^{m,m-1}) = f_2(R_\alpha{}^{21}, \ldots, R_\alpha{}^{m1}) : N_{m-1}\{\mu; \sigma^2 V_{m-1}\}$$

where

$$\mu = \begin{pmatrix} \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_m \end{pmatrix}.$$

The joint density function in the case of an α where $p_{\alpha i^*}$ is missing is given by (16) if $i^* \neq 1$ and (16′) if $i^* = 1$.

(16)
$$f_3[R_\alpha{}^{21}, \ldots) R_\alpha{}^{i^*1}, \ldots, R_\alpha{}^{i,i^*-1}, R_\alpha{}^{i^*+1,i^*}, \ldots, R^{mi^*}(\ldots, R_\alpha{}^{m,m-1}] = f_3[R_\alpha{}^{21}, \ldots) R_\alpha{}^{i^*1}(\ldots, R_\alpha{}^{m1}] : N_{m-2}\{\mu^{)i^*(}; \sigma^2 V_{m-2}\}$$

where

$$\mu^{)i^*(} \equiv \begin{pmatrix} \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu^* \\ \cdot \\ \cdot \\ \cdot \\ \mu_m \end{pmatrix} \qquad i^* = 1$$

(16′)
$$f_4[R_\alpha{}^{32}, \ldots) R_\alpha{}^{21}, \ldots, R_\alpha{}^{m1}(\ldots R_\alpha{}^{m,m-1}] = f_4[R_\alpha{}^{32}, \ldots, R_\alpha{}^{m2}] : N_{m-2}\{\mu_{t2}; \sigma^2 V_{m-2}\}$$

where

$$(\mu_{i2}) \equiv \begin{pmatrix} \mu_{32} \\ \cdot \\ \cdot \\ \cdot \\ \mu_{m2} \end{pmatrix} \quad \text{and} \quad \mu_{i2} = \mu_i - \mu_2 \qquad i^* = 1.$$

The expression for the joint density function of the prices of a given row of P becomes quite complicated notationally when there are many holes. It will be a joint multivariate normal function of $(m - 1 - h_\alpha)$ variables, where h_α is the number of holes in the row, but writing down the general function requires quite awkward notation. The density function will be skipped over here in order to pass directly to the likelihood function of P as a whole. This is given by (17).

$$(17) \quad L(P|\mu_2, \ldots, \mu_m; \sigma^2) = \prod_{\alpha=1}^{A} \frac{1}{(2\pi)^{d/2}} \cdot \frac{1}{\sigma |V_d|^{1/2}} \exp\left[-\frac{1}{2\sigma^2} (\mathscr{R}^{(\alpha)} - M^{(\alpha)})' V_d^{-1} (\mathscr{R}^{(\alpha)} - M^{(\alpha)}) \right]$$

where $\mathscr{R}^{(\alpha)}$ is the $(d \times 1)$ vector of non-redundant R_α^{ij}'s in the α'th row; $M^{(\alpha)}$ is the (d $\times$ 1) vector of $(\mu_i - \mu_j)$ terms associated with the $\mathscr{R}^{(\alpha)}$'s; and $d = m - 1 - h_\alpha$.

The individual factors of L are associated with rows of P. Rows with no holes give rise to $(m - 1)$ dimensional factors, each being a function of $\mu_2, \ldots, \mu_m$, and σ^2. If $p_{\alpha^* 1}$ is missing, then the factor corresponding to the α^* row will still contain all of these arguments; however, if prices of countries other than the base country are missing, the μ_i's corresponding to these countries will not appear in the factor. Thus in general L will be a function of $\mu_2, \ldots, \mu_m$ and σ^2. The set of sufficient statistics for the μ_i's is not obvious. It might be expected that $I_{ij}^{(3)}$, the geometric mean of the available Country i—Country j price ratios, would be the sufficient statistics, but it turns out that instead there is a larger, less simple collection of *weighted* geometric means making up the sufficient set.[4] There is no

[4]The explanation for this is best given by example. Suppose

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ — & p_{22} & p_{23} & p_{24} \\ — & — & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & — \end{pmatrix} \begin{matrix} A = 4 \\ m = 4 \end{matrix}$$

Then

$$L(P|\mu_2, \mu_3, \mu_4; \sigma^2) = \left\{ \frac{1}{(2\pi)^{3/2}} \cdot \frac{1}{\sigma |V_3|^{1/2}} \exp\left[-(1/2\sigma^2) Q_1\right] \right\} \cdot \left\{ \frac{1}{2\pi} \cdot \frac{1}{\sigma |V_2|^{1/2}} \exp\left[-(1/2\sigma^2) Q_2\right] \right\}$$
$$\cdot \left\{ \frac{1}{(2\pi)^{1/2}} \cdot \frac{1}{\sigma} \exp\left[-(1/2\sigma^2) Q_3\right] \right\} \cdot \left\{ \frac{1}{2\pi} \cdot \frac{1}{\sigma |V_2|^{1/2}} \exp\left[-(1/2\sigma^2) Q_4\right] \right\}$$

where

$$Q_1 = \tfrac{3}{2}(R_1{}^{21} - \mu_2)^2 + \tfrac{3}{2}(R_1{}^{31} - \mu_3)^2 + \tfrac{3}{2}(R_1{}^{41} - \mu_4)^2 + (R_1{}^{21} - \mu_2)(R_1{}^{31} - \mu_3)$$
$$- (R_1{}^{21} - \mu_2) \cdot (R_1{}^{41} - \mu_4) - (R_1{}^{31} - \mu_3) \cdot (R^{41} - \mu_4);$$
$$Q_2 = \tfrac{4}{3}(R_2{}^{32} - \mu_3 + \mu_2)^2 + \tfrac{4}{3}(R_2{}^{42} - \mu_4 + \mu_2)^2 - \tfrac{4}{3}(R_2{}^{32} - \mu_3 + \mu_2) \cdot (R_2{}^{42} - \mu_4 + \mu_2);$$
$$Q_3 = (R_3{}^{43} - \mu_4 + \mu_3)^2;$$

[*Continued at foot of next page*

need to spell out the general likelihood function in explicit detail here because in Section IV this stochastic model is approached from a quite different angle. If this alternative way of looking at L were not available, of course the next step would be to find the maximum likelihood estimators of the μ_i's and σ^2. In this case the first-order conditions for a maximum (the equations obtained by setting the various partial derivatives of L equal to zero) give rise to a set of linear equations—but not with a symmetric coefficient matrix—which can be solved directly. Recalling from (7′) that $\mu_i = \ln(p_i^*/p_1^*)$, the invariance property of maximum likelihood estimators assures us that $(p_i^*/p_1^*) = \exp(\hat{\mu}_i)$. The sampling distribution of the $\hat{\mu}_i$'s has not been investigated but clearly the variance-covariance matrix of the $\hat{\mu}_i$'s can be found at least asymptotically from the Hessian of $\ln L$.

Fortunately, the stochastic model developed in this section can be shown to be a special case of the regression model, and therefore its empirical implementation can be carried out with familiar and convenient computing procedures.

IV. A Regression Approach

The fundamental stochastic relationship hypothesized in Section III was given by (8). Provided we take into account (11) and the implications of (12)—the redundancies and the interdependencies of the e_α^{ij}'s—we can rewrite the relationship as in (18).

$$R_\alpha^{ij} = \mu_i - \mu_j + \epsilon_\alpha^{ij}; f(\epsilon_\alpha^{21}, \ldots, \epsilon_\alpha^{m,m-1}): N_{d_\alpha}\{\bar{0}_{d_\alpha}; \sigma^2 V_{d_\alpha}\}. \tag{18}$$

This can be put in the form of a linear regression equation by defining a set of dummy variables, $X_{1\alpha} \ldots, X_{m\alpha}$ such that (18) can be rewritten as (19).

$$R_\alpha^{ij} = \beta_1 X_{1\alpha} + \beta_2 X_{2\alpha} + \ldots + \beta_m X_{m\alpha} + \mu_\alpha^{ij} \tag{19}$$

where

$$X_{k\alpha} = 1 \text{ if } k = i;\ X_{k\alpha} = -1 \text{ if } k = j; \text{ and } X_{k\alpha} = 0 \text{ if } k \neq i, j.$$

[*Continued from foot of previous page*

and

$$Q_4 = \tfrac{4}{3}(R_4^{21} - \mu_2)^2 + \tfrac{4}{3}(R_4^{31} - \mu_3)^2 - \tfrac{4}{3}(R_4^{21} - \mu_2) \cdot (R_4^{31} - \mu_3).$$

The four terms in the braces are derived from the four rows of P.

Now $\ln L$ will be examined.

$$\ln L = -4 \ln(2\pi) - 4 \ln \sigma - \tfrac{1}{2}[\ln |V_3| + 2 \ln |V_2|] - \frac{1}{2\sigma^2}[Q_1 + Q_2 + Q_3 + Q_4].$$

By suitably manipulating the four Q_i's, $\ln L$ can be rewritten in the following form:

$$\begin{aligned}\ln L = g(P, \sigma^2) - \frac{1}{2\sigma^2}\{&\tfrac{17}{6}(\bar{R}_{(1)}^{21} - \mu_2)^2 + \tfrac{17}{6}(\bar{R}_{(2)}^{31} - \mu_3)^2 + \tfrac{8}{2}(\bar{R}_{(3)}^{41} - \mu_4)^2 \\ &+ \tfrac{4}{3}(\bar{R}_{(4)}^{32} - \mu_3 + \mu_2)^2 + \tfrac{4}{3}(\bar{R}_{(5)}^{42} - \mu_4 + \mu_2)^2 + (\bar{R}_{(6)}^{43} - \mu_4 + \mu_3)^2 \\ &- \tfrac{7}{3}(\bar{R}_{(7)}^{21} - \mu_2) \cdot (\bar{R}_{(8)}^{31} - \mu_3) - (\bar{R}_{(9)}^{21} - \mu_2) \cdot (\bar{R}_{(10)}^{41} - \mu_4) - (\bar{R}_{(11)}^{31} - \mu_3) \\ &\times (\bar{R}_{(12)}^{41} - \mu_4) - \tfrac{4}{3}(\bar{R}_{(13)}^{32} - \mu_3 + \mu_2) \cdot (\bar{R}_{(14)}^{42} - \mu_4 - \mu_2)\}\end{aligned}$$

where $\bar{R}_{(n)}^{ij}$ is a weighted average of the logs of all available $p_{\alpha i}/p_{\alpha j}$ ratios. The weights vary for different $i-j$ comparisons; and they also differ for different terms. For example,

$$\bar{R}_{(1)}^{21} = (\tfrac{3}{2} R_1^{21} + \tfrac{4}{3} R_4^{21})/\tfrac{17}{6}, \text{ but } \bar{R}_{(7)}^{21} = (\tfrac{1}{2} R_1^{21} + \tfrac{3}{8} R_4^{21})/\tfrac{7}{6}. \text{ The } \bar{R}_{(n)}^{ij}$$

constitute a set of sufficient statistics for μ_2, μ_3, and μ_4.

The observations that (19) refers to are the set R_α^{i1} $(i > 1)$ for all rows in which $p_{\alpha 1}$ is present, R_α^{i2} $(i > 2)$ for all rows in which $p_{\alpha 1}$ is missing but $p_{\alpha 2}$ is present, and in general R_α^{ik} $(i > k)$ for all rows in which $p_{\alpha 1}, p_{\alpha 2}, \ldots, p_{\alpha,(k-1)}$ are all missing but $p_{\alpha k}$ is present. The regression equation departs from the classical linear model only because the variance-covariance matrix of the disturbances is not diagonal. In fact, the variance-covariance matrix, Ω, consists of blocks V_{d_α} running down the principal diagonal as given in (20).

$$\Omega = \sigma^2 \begin{pmatrix} V_{d_1} & 0 & \ldots & 0 \\ 0 & V_d & \ldots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \ldots & V_{d_A} \end{pmatrix}. \tag{20}$$

Since V_{d_α} consists of known elements—ones down the principal diagonal and $\frac{1}{2}$'s elsewhere—Aitken's method can easily be employed to find the maximum likelihood estimates of the β_i's of (19).

Of course, then the variance-covariance matrix of the $\hat{\beta}_i$'s can be obtained easily, because it is simply $\hat{\sigma}^2 (X'\Omega^{-1}X)^{-1}$, where X is the matrix of observations on the dummy variables $X_1, \ldots, X_m$.

The price-level comparison between Country i and Country j is given in (19) by $(\beta_i - \beta_j)$. Then the maximum-likelihood estimate of the relative price levels of the two countries is $(\hat{\beta}_i - \hat{\beta}_j)$ from the empirical regression based upon the nonredundant R_α^{ij}'s, and the standard error is given by $\sqrt{\hat{\sigma}_{\hat{\beta}_i}^2 + \hat{\sigma}_{\hat{\beta}_j}^2 - 2\hat{\sigma}_{\hat{\beta}_i\hat{\beta}_j}}$ where the individual elements come out of the variance-covariance matrix of the $\hat{\beta}_i$'s.

The regression of (19) is well worth settling for as a way of attacking the likelihood function of (17). However, it turns out that we can do better than (19). The history of price index construction has been shaped by the fact that what is interesting about price-levels—for time-to-time or place-to-place comparisons—is the *relative* level of one set of prices to another. As a consequence, price-ratios were the natural element of observation. The fact that price-ratios are units-invariant seemed to add additional weight to the argument that price-ratios or some function thereof should be the basic dependent variable. But working with price-ratios makes the regression based upon (19) unnecessarily complicated. The fact that the non-redundant disturbance terms are not all mutually independent springs from the appearance of the same common base country price in the denominator of all the price ratios associated with any particular row of P. Suppose as an alternative to (6) and (18), we stipulate (21) and (21′).

$$p_{i\alpha} = p_i^* \cdot \bar{w}_\alpha^i; f(\bar{w}_\alpha^1, \ldots, \bar{w}_\alpha^m): \quad \text{Lognormal } [\bar{0}, \sigma^2 I] \tag{21}$$

$$\ln p_{i\alpha} = \bar{\mu}_i + \delta_\alpha + v_\alpha^i; f(v_\alpha^1, \ldots, v_\alpha^m): N_m\{\bar{0}_m; \sigma^2 I\} \tag{21′}$$

where $\bar{\mu}_i$ is the natural log of the i'th country's price-level.

Then (22), the linear regression equation springing from (21′), involves two sets of dummy variables $(X_{i\alpha}, Y_{i\alpha})$, $i = 2, \ldots, m$; $\alpha = 1, \ldots A$.

$$\ln p_{\alpha i} = \beta_2 X_{2\alpha} + \ldots + \beta_m X_{m\alpha} + \gamma_1 Y_{i1} + \ldots + \gamma_A Y_{iA} + v_\alpha^i \tag{22}$$

The advantage of (22) over (19) is that now the disturbance term meets the standard conditions for classical regression. (Incidentally, observe that the assumption of lognormality of the underlying disturbance terms, either w_α^{ij} or w_α^1, is not essential; without the assumption one can still rely on the optimal properties flowing from the Gauss–Markov theorem on Least Squares.)

The coefficient of the $X_{i\alpha}$ dummy, β_i, in (22) is to be interpreted as the natural log of the ratio of the price-level in Country i to the price-level in Country 1, the base country. Thus $\exp(\tilde{\beta}_i)$ is an estimate of p_i^*/p_1^*. (Under the assumption of lognormality of the original disturbances, $\exp(\tilde{\beta}_i)$ will be the maximum likelihood estimate of p_i^*/p_1^*, but the expected value of $\exp(\tilde{\beta}_i)$ is not p_i^*/p_1^*. Similarly, $\exp(\tilde{\beta}_i - \tilde{\beta}_j)$ is an estimate of p_i^*/p_j^*. Clearly, $\exp(\tilde{\beta}_i - \beta_j) = 1/\exp(\tilde{\beta}_j - \tilde{\beta}_i)$ and $\exp(\tilde{\beta}_i - \tilde{\beta}_j) = \exp(\tilde{\beta}_i - \tilde{\beta}_k)/\exp(\tilde{\beta}_j - \tilde{\beta}_k)$. Therefore, the estimates of relative price-levels are both base invariant and circular.

The coefficients of the γ_α dummies have no significance for country price-level comparisons, though they may conceivably be useful in an analysis of the relative values that purchasers in all m countries put on the individual items. (If, for example, the units of the items are standardized for an attribute—say, caloric content in the case of a fuel category or nutritional content in the case of a food category—so that a unit of each item contains the same quantity of the attribute, then $(\gamma_{\alpha_1} - \gamma_{\alpha_2})$ is directly relevant to the question of how attractive Item 1 is compared with Item 2 as a way of securing a standard unit of the attribute. Specifically, if $\gamma_7 - \gamma_3 = 0.21$, then the cost of obtaining a unit of the attribute through the purchase of the seventh item would be 23 per cent ($e^{0.21} = 1.23$) greater than if it were obtained through the purchase of the third item). Since a change in the units of an item will lead only to a change in the Y-dummy associated with the item, the price-level estimates for the various countries will be invariant under a change in units.

The connection between the regressions of (22) and (19) can be made clearer with an analogy. Suppose it is believed that household consumption is related to household income by a linear consumption function. Using a set of data, (C_i, Y_i), a conventional regression could be run to estimate the marginal propensity-to-consume. Alternatively, one could single out a particular household, say the Jones family, and run a regression with a suppressed constant on the data set $\{C_i - C_{\text{Jones}}, Y_i - Y_{\text{Jones}}\}$. This latter regression would not be a very satisfactory one because it would be based upon one less observation than the other and because the interdependence of its disturbance terms $[E(u_i - u_{\text{Jones}}) \cdot (u_j - u_{\text{Jones}}) = \sigma^2;\ E(u_i - u_{\text{Jones}})^2 = 2\sigma^2]$ would call for the use of Aitken's method. Unless the Jones family was particularly important in the analysis of consumer spending behavior, there would be strong reasons for avoiding assigning to it the asymmetric role implied by the $(C_i - C_{\text{Jones}} : Y_i - Y_{\text{Jones}})$ regression. Similarly, if the base country has been selected only because it provides a convenient numeraire, it would be better to use the regression of (22) than the regression of (19)[5].

[5]The omission of an X-dummy for the base country does not assign the base country a special position. Literally nothing substantive would be changed if $\beta_1 X_{1\alpha}$ was added and $\gamma_1 Y_{1\alpha}$ was dropped from (22).

Before passing on to an empirical example, a last set of comments should be made about the precision of the price-level estimates and the notion of randomness. First, the regression procedure delivers estimates of the standard errors of the regression coefficient estimates. These of course provide the basis for computing confidence intervals for the true regression coefficients. Since the price-level ratios are simply the exponentials of the regression coefficients, it is an easy matter to go on to compute confidence intervals for the price-level ratios themselves. Specifically.

$$(23) \qquad P\{\exp(\tilde{\beta}_i - t_\pi \tilde{\sigma}_{\tilde{\beta}_i}) < \frac{p_i^*}{p_1^*} < \exp(\tilde{\beta}_i + t_\pi \tilde{\sigma}_{\tilde{\beta}_i}\} = \pi$$

where t_π is an appropriate entry from a Student's t distribution table. (Qualification: (23) holds strictly only if the disturbance terms of (22)—the $v_\alpha{}^i$'s—are normal. If they are not, the Student's t distribution is not the correct source for t_π. However, if the number of degrees of freedom of the regression is large—i.e., the number of prices actually present in the P tableau minus the number of parameters estimated in the regression, $A + (m - 1)$, is greater than, say, thirty, (23) is likely to be an acceptable approximation to a "true" confidence interval.)

Secondly, the notion of random sampling needs amplification. In regression analysis, it is the disturbances which must be randomly distributed, not the independent variables. Thus, it is not necessary that the holes in P be randomly distributed, provided that the systematic pattern of the holes—with respect to countries and items—does not lead to a systematic pattern among the disturbances.[6] The fact that countries at different stages of development may not consume identical goods and services within a category may give rise to a nonrandom pattern of holes, but this does not necessarily introduce any bias in the regression coefficient estimates. The only concern about the pattern of the holes is that they do not lead to a singular variance-covariance matrix for the independent variables. (Singularity could occur if, to give one example, the set of goods in a detailed category and the set of countries each can be divided into two subsets such that no member of the first goods subset is priced in the first country subset, and no member of the second goods subset is priced in the second country subset. Intuitively, this simply means that country price levels cannot be compared unless there is some overlap in the list of goods which have been priced in the two countries.) To summarize: the random sampling requirement is simply that the price of an item in a country should depart from an amount defined by the country's price level and the units of the item only by an amount which stochastically does not depend upon either the country or the item.

[6]Suppose my research assistant prepared the tableau P but before a regression was run on the whole set of data he splattered opaque coffee on the data sheet in such a way that a "random" set of prices was obliterated. The regression run on the still-legible prices would be less satisfactory than one based upon all of the data, of course, but it still would give unbiased estimates of the price-levels. However, if he for some reason were angry and wanted to do me harm, he could cause me real trouble. Suppose he blotted out the same number of prices but on a systematic basis: low-price points in some arrays and high-price points in others. By appropriately partitioning the data set and running twin regressions, it would be possible to determine (in probabilistic terms) whether the missing observations are a product of mischief rather than carelessness. Analogously, it would be possible to tell if the missing observations resulting from failure to price all items leads to bias in estimating relative price-levels.

V. Example: Food

To illustrate these ideas, regressions based upon (22) have been carried out on the hypothetical price tableau given in Table 1. The category, Food, is in fact much, much broader than those envisioned for the International Comparison Project,[7] but the data are still suitable for this exercise.

TABLE 1

Prices of Nine Food Items in Each of Seven Countries in 1967

Item		Countries						
		U.S. ($) X_1	Germany (DM) X_2	Japan (yen) X_3	U,K, (shilling) X_4	Kenya (shilling) X_5	Colombia (peso) X_6	India (rupee) X_7
Eggs	(kg) Y_1	0.49	2.64	—	3.96	4.38	9.74	3.00
Milk	(kg) Y_2	0.27	—	483.12	1.71	1.44	—	1.08
Butter	(kg) Y_3	1.83	7.82	777.77	7.39	8.36	28.63	—
Oranges	(kg) Y_4	0.32	1.50	118.00	2.79	0.99	—	—
Bread	(kg) Y_5	0.49	—	110.00	2.18	1.65	8.88	—
Potatoes	(kg) Y_6	0.16	—	51.20	0.70	0.46	1.46	0.77
Sugar	(kg) Y_7	0.27	—	133.00	1.54	0.70	2.35	3.63
Bacon	(kg) Y_8	1.69	6.69	—	—	14.90	32.22	—
Cabbage	(kg) Y_9	0.24	—	38.30	1.14	0.79	—	—

—: Price not available.

The regression result for the nine items and seven countries (after all national prices were converted to American dollars at the official exchange rates prevailing in 1967) is given in (23).

$$\begin{aligned}(24)\quad \ln p_{\alpha i} = &- \underset{(0.257)}{0.008}\, X_2 - \underset{(0.209)}{0.173}\, X_3 - \underset{(0.200)}{0.214}\, X_4 - \underset{(0.193)}{0.510}\, X_5 - \underset{(0.221)}{0.113}\, X_6 \\ &- \underset{(0.256)}{0.252}\, X_7 - \underset{(0.213)}{0.420}\, Y_1 - \underset{(0.224)}{1.039}\, Y_2 + \underset{(0.211)}{0.583}\, Y_3 - \underset{(0.223)}{1.120}\, Y_4 \\ &- \underset{(0.222)}{0.894}\, Y_5 - \underset{(0.212)}{2.101}\, Y_6 - \underset{(0.212)}{1.315}\, Y_7 + \underset{(0.242)}{0.780}\, Y_8 - \underset{(0.238)}{1.790}\, Y_9\end{aligned}$$

The numbers in parentheses are the standard errors of the coefficients directly above them. These coefficients give:

$$\left(\frac{\tilde{p_2}^*}{p_1^*}\right) = 0.992;\ \left(\frac{\tilde{p_3}^*}{p_1^*}\right) = 1.189;\ \left(\frac{\tilde{p_4}^*}{p_1^*}\right) = 0.808;$$

$$\left(\frac{\tilde{p_5}^*}{p_1^*}\right) = 0.601;\ \left(\frac{\tilde{p_6}^*}{p_1^*}\right) = 0.893;\ \text{and}\ \left(\frac{\tilde{p_7}^*}{p_1^*}\right) = 0.778,$$

[7] In fact, there are 36 individual detailed categories comprising Food in the International Comparison Project.

where the numerator subscript is keyed to the numbers of the countries as ordered in Table 1.

Table 2 gives a somewhat more general, but confirming picture of the relationship of the countries to the United States. In order to see how sensitive the price-level estimates are to exactly what items and which countries are represented in the data, repeat regressions were carried out with various omissions of rows and columns of Table 1. The results were:

(1) Regressions based upon all countries but upon either nine or seven or eight items all gave pretty much the same relative price estimates. The exclusion of cabbage from the regression, a relatively cheap item in Japan, increased that country's estimated relative price level by about 15 percent, but no other discrepancy came close to that magnitude.

(2) Regressions based upon all nine items but varying the number of countries gave very similar relative price-level estimates. The only exception was India where dropping three other countries led to a 17 percent increase in the estimate of India's price-level relative to the United States.

TABLE 2

THE PRICE-LEVELS RELATIVE TO THE UNITED STATES OF EACH OF SIX COUNTRIES AS ESTIMATED IN DIFFERENT WAYS*

Estimation Method Using (22) Items	Countries	Germany	Japan	U.K.	Kenya	Colombia	India
1, 2, 3, 4, 5, 6, 7, 8, 9	1, 2, 3, 4, 5, 6, 7	0.992 [5]	1.189 [6]	0.808 [3]	0.601 [1]	0.893 [4]	0.778 [2]
1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 5, 6, 7	1.004 [5]	1.343 [6]	0.801 [3]	0.563 [1]	0.907 [4]	0.785 [2]
1, 2, 3 ,4, 5, 6, 7, 8	1, 2, 3, 4, 5, 6, 7	1.032 [5]	1.389 [6]	0.827 [3]	0.621 [1]	0.933 [4]	0.814 [2]
1, 2, 3, 4, 5, 6, 7, 8, 9	1, 2, 3, 4, 5	0.988 [3]	1.195 [4]	0.807 [2]	0.601 [1]	—	—
1, 2, 3, 4, 5, 6, 7, 8, 9	1, 2, 6, 7	1.050 [3]	—	—	—	0.872 [1]	0.907 [2]
1, 2, 3, 4, 5, 6, 7, 8, 9	1, 3, 4, 5	—	1.217 [3]	0.819 [2]	0.601 [1]	—	—
*Geometric Mean of Available Price Ratios***		1.139(4) [6]	1.104(7) [5]	0.782(8) [2]	0.600(9) [1]	0.903(6) [4]	0.836(4) [3]

*The rankings of the countries appear in brackets below each relative price-level.
**The number of available price-ratios appears in round parentheses.

The last row of Table 2 gives the estimates of relative price levels obtained by computing the geometric means of all available price-ratios. For three countries the geometric mean estimates are virtually the same as the estimates from the "full-information" regression for nine items and seven countries. But the German and Indian estimates are 15 and 8 percent higher respectively and the Japanese estimate is 8 percent lower. Notice that dropping observations from the

regressions did not change the ranking of the six countries, though the position of the United States relative to Germany shifted slightly from regression to regression. The ranking obtained from the geometric means was different: the positions of India and the UK were switched, as were the positions of Germany and Japan. (It is no coincidence that these switches involved countries for which only four prices were available). Incidentally, the estimated price index for Germany relative to Colombia was 1.110 as derived from the regression method but only 1.078 using the geometric method (based upon three price ratios).

TABLE 3

0.95 CONFIDENCE INTERVALS FOR ESTIMATES OF RELATIVE PRICE-LEVELS

	Regression*		Geometric Mean**	
	Lower	Upper	Lower	Upper
Germany/U.S.	0.59	1.67	0.92	1.41
Japan/U.S.	0.78	1.82	0.44	2.75
U.K./U.S.	0.58	1.13	0.44	1.39
Kenya/U.S.	0.41	0.89	0.42	0.86
Colombia/U.S.	0.57	1.40	0.62	1.31
India/U.S.	0.46	1.31	0.36	1.95

*Based upon data of 9 items and 7 countries.
**Based upon all available price ratios.

The confidence intervals for relative price-levels as obtained from the regressions appear to be distressingly wide. This is partly because of the heterogeneity of the food prices of Table 1 and partly because the regression estimates 15 parameters on the basis of only 47 observations. The 0.95 confidence intervals are given in Table 3. In addition, 0.95 confidence intervals based upon the geometric means of available price-ratios are also given. It was expected that the geometric mean intervals would be wider than the corresponding regression ones. The reasoning behind this was (1) the regression approach, taking advantage of circularity, uses the data more efficiently, and (2) the very small numbers of price-ratios on which the geometric means are based lead to relatively large t values in the confidence intervals formula. In fact, half of the geometric mean C.I.'s are smaller than the regression ones but in two of the cases, Kenya and Colombia, the superiority is only marginal. The regression C.I.'s are smaller in three cases and in two of these, Japan and India, the superiority is substantial. This suggests the conjecture that the *a priori* argument that the regression C.I.'s are smaller perhaps should be modified to include the qualifier "on the average". (It should be remarked that the C.I.'s derived from the geometric means are much less likely to be robust with respect to the assumption of lognormality than those derived from the regression.)

VI. CONCLUDING REMARKS

We have examined in detail the question of how to estimate, on the basis of incomplete price data, relative price-levels of a number of different countries for a

narrow category of output. Happily, the answer turns out to be a simple one. The most commonly used technique in empirical economics, regression analysis, can be harnessed to do the task with only a minimum of complications in transforming the price data into a regression format. In the limiting case, when there are no missing observations in the price tableau, it can be shown that the regression procedure amounts to the computation of a set of geometric means. Since this is just what one normally would compute to estimate relative price-levels if he did not introduce stochastic considerations into his framework of analysis, it is reassuring to see that the regression method is consistent with ordinary practice. Estimating precision is always important so casting the problem in stochastic terms is useful even in the complete-data case.[8]

In closing the reader is reminded that no effort has been directed at the greatest difficulty of all, the quantity-weighting problem. However, when the definitive empirical work on the index number problem has been completed, a stochastic model of the general sort worked out above will surely be at stage-center.

[8]There is another advantage of the regression approach. The fact that we know a great deal about regression analysis makes modifications easier. Suppose, for example, that the individual prices of P are outputs from a set of hedonic index regressions and therefore themselves should be regarded as having standard errors which depend upon the individual hedonic regressions. If these standard errors are not the same for all prices, then it would be hard to know just how to weight the observations in the conventional procedure. However, weighted regressions are well-understood.