Review of Income and Wealth
Series 49, Number 1, March 2003

ON THE VARIANCE/COVARIANCE STRUCTURE OF THE LOG FISHER INDEX, AND IMPLICATIONS FOR AGGREGATION TECHNIQUES

BY JAMES R. CUTHBERT*

Edinburgh, U.K.

Several aggregation methods, including the EKS, start by calculating bilateral Fisher indices. Prices and quantities are, however, subject to measurement error. This stochastic behavior, which implies both unequal variances, and non-zero correlations, between different Fisher indices, has to be taken into account if optimal estimates of aggregate PPPs are to be derived from the Fisher indices. This paper provides estimates of the variance/covariance structure of the Fisher indices, under two alternative models for stochastic variation at basic heading level: and it applies these formulae to the 1996 OECD data set, illustrating that the Fisher indices for this data set are indeed highly correlated. The paper also establishes a general theoretical result, proving that the EKS is optimal for a particular variance/covariance structure involving non-zero correlations, and hence shows that the standard EKS aggregation method is likely to be near optimal for the 1996 OECD data set.

1. INTRODUCTION

The aggregation problem is concerned with how to produce estimates of comparative real GDPs, given data on prices and quantities at individual item or commodity level. This paper is concerned with multilateral aggregation techniques which start from the bilateral Fisher index. In particular, recognizing that the individual item level price and quantity observations are subject to errors of observation, it assesses what the implications of these errors are for the variance and covariance structure of the bilateral Fisher indices: it then goes on to examine the implications of this variance/covariance structure for aggregation techniques based on the Fisher index.

The notation used in this paper is that there are I items, and J countries, and that

$$p_{ij} = \text{price of item } i \text{ in country } j;$$
$$q_{ij} = \text{quantity of item } i \text{ in country } j.$$

The bilateral Fisher volume index, denoted here by F_{jk}, is defined as the geometric mean of the Laspeyres and Paasche volume indices, L_{jk} and P_{jk}, so:

$$F_{jk} = [L_{jk} P_{jk}]^{0.5},$$

*Correspondence to: James Cuthbert, 42 Cluny Drive, Morningside, Edinburgh EH10 6DX, U.K. (jamcuthbert@blueyonder.co.uk).

where:

$$L_{jk} = \frac{\sum_i p_{ik} q_{ij}}{\sum_i p_{ik} q_{ik}},$$

and

$$P_{jk} = \frac{\sum_i p_{ij} q_{ij}}{\sum_i p_{ij} q_{ik}}.$$

The Fisher index is often regarded as having ideal properties in bilateral comparisons: (see also Balk (1995), for a discussion of the Fisher index in relation to axiomatic price theory). For multilateral comparison purposes, however, straightforward application of the Fisher index is not possible, since the Fisher index does not satisfy the important criterion of transitivity: that is, in general,

$$F_{jk} \neq F_{jm} F_{mk}.$$

One general approach to the aggregation problem is to attempt to approximate the structure of the bilateral set of Fisher indices by means of a transitive (that is, multiplicative), model. Equivalently, if the logarithms of the bilateral Fisher indices are taken, then the log Fisher indices can be regarded as being determined in terms of a linear model, of the following form,

$$y_{jk} = \log(F_{jk}) = c_j - c_k + \varepsilon_{jk}, \quad k > j \tag{1}$$

where the c_j are unknown parameters, and the ε_{jk} are unknown error terms. The parameters in model (1) can be estimated by regression techniques, giving estimates $\hat{c}_j$, say: the required multiplicative approximation to the structure of the original Fisher indices is then given by

$$\exp(\hat{c}_j - \hat{c}_k), \quad j = 1 \ldots J, k = 1 \ldots J.$$

If the errors ε_{jk} in equation (1) are uncorrelated, and of equal variance, then the estimation of the parameters c_j by standard regression gives estimators $\hat{c}_j$ satisfying

$$\hat{c}_j - \hat{c}_k = \frac{1}{J} \sum_n (y_{jn} - y_{kn}),$$

where this equation is to be interpreted in terms of the convention that $y_{jj} = 0$, and $y_{kj} = -y_{jk}$, $k > j$. (This result is well known, and is easy to prove from first principles, by setting up the normal equations for the standard regression estimator, and then verifying that the estimator $\hat{c}_j$ satisfies these equations.) The estimator is equivalent to the well known EKS estimator, defined as

$$\mathrm{EKS}_{jk} = \left[\prod_n F_{jn} F_{nk} \right]^{\frac{1}{J}}.$$

The EKS (or Elteto Koves Szulc) method of aggregation was developed by Elteto and Koves (1964) and Szulc (1964); the method was independently anticipated by

Gini (1924), and is therefore sometimes referred to as the GEKS method. See Hill (1997), for a discussion of the EKS in relation to a taxonomy of aggregation methods.

If the errors ε_{jk} are written in the form of a vector **e** of length $\frac{(J-1)J}{2}$, then another way of expressing the assumption that the ε_{jk} terms are uncorrelated and of equal variance is to say that the variance/covariance matrix of this vector is proportional to the identity matrix: i.e. $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$, where σ^2 is an unknown variance parameter, and **I** is the identity matrix. However, in practice, the error terms ε_{jk} in equation (1) are neither likely to be of equal variance, nor to be uncorrelated: in other words, in general,

$$\text{var}(\mathbf{e}) = \sigma^2\mathbf{E}, \tag{2}$$

where **E** is some positive definite symmetric matrix, with unequal diagonal terms (the variance terms), and at least some non-zero terms in off-diagonal positions (i.e. at least some non-zero correlations between error terms).

The problem then arises, of how to estimate the parameters in equation (1), when the variance/covariance matrix of the error terms has the more general form in equation (2). The theoretical answer to this problem, when the matrix **E** is known, is given by the technique of Generalized Least Squares (GLS) (see Johnston and Dinardo, 1997). GLS theory states that, for a linear model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, where $\text{var}(\mathbf{e}) = \sigma^2\mathbf{E}$, then the best linear unbiased estimator of the parameter vector **b** is given by the GLS estimator

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{E}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}^{-1}\mathbf{y};$$

equivalently, if $\mathbf{E}^{-1}$ is expressed in terms of a matrix **W** as $\mathbf{E}^{-1} = \mathbf{W}'\mathbf{W}$, as can always be done, then **b** can be estimated by the application of Ordinary Least Squares to the transformed model

$$\mathbf{Wy} = \mathbf{WXb} + \mathbf{We}.$$

In the special case where **E** is a diagonal matrix (which corresponds to uncorrelated errors in the original model), then this last formula corresponds to the familiar application of weighted regression, with weights inversely proportional to the standard errors of the error terms in the original model.

While GLS supplies the appropriate theory for optimal estimation of the coefficients in equation (1), two problems remain. Firstly, there is the problem of determining the correct form of the variance/covariance matrix **E**. Secondly, there is the problem of deriving the GLS estimator, if **E** has a complicated form. This paper is concerned with both of these problems. Firstly, the theoretical form of the variance/covariance matrix of the errors in equation (1) is derived: this is done for two specific models, each specifying a different assumption about the form of stochastic variation at the level of the individual item. Second, the actual variance/covariance structure of the log Fisher indices is estimated, for each of these models, using the 1996 OECD comparison data set. It is shown that the assumption of uncorrelated errors is very far from holding. A general theoretical result is also established, proving that the EKS is optimal for a particular variance/covariance structure involving non-zero correlations. In the light of this

result, it is shown that the conventional EKS estimator is still likely to be near optimal for the OECD data set.

Historically, an early stochastic formulation for the variance of the log Fisher index, together with the use of these variances as weights in a weighted log Fisher regression, was given by Cuthbert and Cuthbert (1988). This, however, was for the degenerate form of the Fisher index used in aggregating up to basic heading level; nor did this work take into account the covariances between the log Fisher indices.

More recently, Rao (2001) described a weighted EKS approach, both for aggregation below basic heading level, and from basic heading up. Rao's approach implicitly assumes that the variance/covariance structure of the log Fisher terms is diagonal. He discusses a number of possible approaches to the derivation of the appropriate weights, including measures based on the spread between the Paasche and Laspeyres indices, on the economic distance between countries, and on measures of similarity in price structure. These approaches suffer, however, from not being directly related to a model of how the variance of the log Fisher relates to the basic stochastic structure of the data, as well as not taking into account the off-diagonal correlations of the log Fisher indices.

The approach taken in this paper is primarily based on a stochastic approach to index numbers: that is, a model based approach which takes into account the implications of various forms of measurement error or uncertainty. Other approaches, like the axiomatic approach, or the economic approach are possible: see, for example, Balk (2001), for a recent discussion of different possible approaches. See also Diewert (1995), for a critique of the stochastic approach (though a somewhat more narrowly based stochastic approach than that considered here).

In the real world, consideration of the random elements in the estimation of purchasing power parities seems inescapable; items will normally be sampled, and measurement of prices and expenditures will be subject to random measurement errors. In the view of the author, therefore, it is important to consider the stochastic properties of any given index number formulation: this is what this paper is concerned with. This, however, is not inconsistent with the idea that the ultimate rationale underlying a particular formula may be based on considerations originating in the axiomatic or economic approaches to index numbers. In other words, in the author's view the stochastic approach is complementary to the axiomatic and economic approaches, rather than being in opposition to these approaches.

2. The Variance and Covariance Structure of the Log Fisher Index

The individual price and quantity observations p_{ij} and q_{ij} are, of course, subject to various forms of measurement and estimation error; in this paper, two basic models are developed to describe the stochastic structure of these observations.

Under Model 1, it is assumed that the starting point is independent observations of prices and quantities, and that the observed p_{ij} and q_{ij} are related to underlying true values, $\dot{p}_{ij}$ and $\dot{q}_{ij}$, by the following relationships:

$$p_{ij} = \dot{p}_{ij}(1+\phi_{ij}), \quad \text{and} \quad q_{ij} = \dot{q}_{ij}(1+\gamma_{ij}),$$

where the errors of observation ϕ_{ij} and γ_{ij} are independent, and $\text{var}(\phi_{ij}) = \sigma_P^2$, $\text{var}(\gamma_{ij}) = \sigma_Q^2$.

This model describes a situation where, for each country, 95 percent of the price observations are within $\pm 200\sigma_P$ percent of the true underlying value, and 95 percent of the quantity observations are within $\pm 200\sigma_Q$ percent of the true underlying value. The variance parameters σ_P^2 and σ_Q^2 are assumed to be unknown.

Model 1, as described in the previous paragraph, represents probably the simplest model possible for the stochastic structure of the individual observations. However, the model 1 situation, where the price and quantity data are estimated independently, often does not hold in the real world. What often happens in practice is that it is prices and expenditures which are estimated directly for each item, and quantities are derived by dividing expenditures by prices. Model 2 describes this situation. If expenditure on item i in country j is denoted by e_{ij}, then under model 2 it is assumed that the observed p_{ij} and e_{ij} are related to underlying true values, $\dot{p}_{ij}$ and $\dot{e}_{ij}$, by the following relationships:

$$p_{ij} = \dot{p}_{ij}(1+\phi_{ij}), \quad \text{and} \quad e_{ij} = \dot{e}_{ij}(1+\lambda_{ij}),$$

where the errors of observation ϕ_{ij} and λ_{ij} are independent, and $\text{var}(\phi_{ij}) = \sigma_P^2$, $\text{var}(\lambda_{ij}) = \sigma_E^2$.

Under either of the above models, the individual Fisher index terms F_{jk} have a stochastic structure; and the variance/covariance structure of the F_{jk} can be derived theoretically.

As before, let $y_{jk} = \log(F_{jk})$; the variance and covariance terms involving y_{jk} are denoted by

$$v(j,k) = \text{var}(y_{jk}), \text{and}$$
$$\text{co}(j,m,n) = \text{cov}(y_{jm}, y_{jn}).$$

Then approximations to these variance and covariance terms are given by the following formulae, under model 1 and model 2 respectively:

Model 1

(3) $v(j,k) =$

$$\frac{\left(\sigma_P^2 + \sigma_Q^2 + \sigma_P^2\sigma_Q^2\right)}{4}\left[\frac{\sum_i p_{ij}^2 q_{ij}^2}{\left(\sum_i p_{ij}q_{ij}\right)^2} + \frac{\sum_i p_{ik}^2 q_{ik}^2}{\left(\sum_i p_{ik}q_{ik}\right)^2} + \frac{\sum_i p_{ij}^2 q_{ik}^2}{\left(\sum_i p_{ij}q_{ik}\right)^2} + \frac{\sum_i p_{ik}^2 q_{ij}^2}{\left(\sum_i p_{ik}q_{ij}\right)^2}\right]$$
$$+\frac{\sigma_Q^2}{2}\left[\frac{\sum_i p_{ij}p_{ik}q_{ij}^2}{\sum_i p_{ij}q_{ij}\sum_i p_{ik}q_{ij}} + \frac{\sum_i p_{ij}p_{ik}q_{ik}^2}{\sum_i p_{ij}q_{ik}\sum_i p_{ik}q_{ik}}\right]$$
$$-\frac{\sigma_P^2}{2}\left[\frac{\sum_i p_{ij}^2 q_{ij}q_{ik}}{\sum_i p_{ij}q_{ij}\sum_i p_{ij}q_{ik}} + \frac{\sum_i p_{ik}^2 q_{ij}q_{ik}}{\sum_i p_{ik}q_{ij}\sum_i p_{ik}q_{ik}}\right]$$

and

(4) $$co(j,m,n) = \frac{(\sigma_P^2 + \sigma_Q^2 + \sigma_P^2\sigma_Q^2)\sum_i p_{ij}^2 q_{ij}^2}{4\left(\sum_i p_{ij}q_{ij}\right)^2} + \frac{\sigma_Q^2}{4}\left[\frac{\sum_i p_{ij}p_{im}q_{ij}^2}{\sum_i p_{ij}q_{ij}\sum_i p_{im}q_{ij}} + \frac{\sum_i p_{ij}p_{in}q_{ij}^2}{\sum_i p_{ij}q_{ij}\sum_i p_{in}q_{ij}} + \frac{\sum_i p_{im}p_{in}q_{ij}^2}{\sum_i p_{im}q_{ij}\sum_i p_{in}q_{ij}}\right] + \frac{\sigma_P^2}{4}\left[\frac{\sum_i p_{ij}^2 q_{im}q_{in}}{\sum_i p_{ij}q_{im}\sum_i p_{ij}q_{in}} - \frac{\sum_i p_{ij}^2 q_{ij}q_{im}}{\sum_i p_{ij}q_{ij}\sum_i p_{ij}q_{im}} - \frac{\sum_i p_{ij}^2 q_{ij}q_{in}}{\sum_i p_{ij}q_{ij}\sum_i p_{ij}q_{in}}\right]$$

Formula (3) holds for $j \neq k$; note that $v(j, j) = 0$.
Formula (4) holds when j, m and n are all different; note that

$$\text{co}(j,m,m) = v(j,m); \text{ and co}(j,j,m) = \text{co}(j,m,j) = 0.$$

Finally, note that cov(log(F_{jk}), log(F_{mn})) = 0 for j, k, m and n all different; that is, the full covariance matrix of the log(F_{jk}) terms contains large blocks which are structurally zero.

Model 2

(5) $$v(j,k) = \frac{\sigma_E^2}{4}\left[\frac{\sum_i p_{ij}^2 q_{ij}^2}{\left(\sum_i p_{ij}q_{ij}\right)^2} + \frac{\sum_i p_{ik}^2 q_{ik}^2}{\left(\sum_i p_{ik}q_{ik}\right)^2} + \frac{2\sum_i p_{ij}p_{ik}q_{ij}^2}{\sum_i p_{ij}q_{ij}\sum_i p_{ik}q_{ij}} + \frac{2\sum_i p_{ij}p_{ik}q_{ik}^2}{\sum_i p_{ij}q_{ik}\sum_i p_{ik}q_{ik}}\right] + \frac{\left[\left(1+\sigma_P^2\right)^2\left(1+\sigma_E^2\right)-1\right]}{4}\left[\frac{\sum_i p_{ij}^2 q_{ik}^2}{\left(\sum_i p_{ij}q_{ik}\right)^2} + \frac{\sum_i p_{ik}^2 q_{ij}^2}{\left(\sum_i p_{ik}q_{ij}\right)^2}\right]$$

(6) $$\text{co}(j,m,n) = \frac{\sigma_E^2}{4}\left[\frac{\sum_i p_{ij}^2 q_{ij}^2}{\left(\sum_i p_{ij}q_{ij}\right)^2} + \frac{\sum_i p_{ij}p_{im}q_{ij}^2}{\sum_i p_{ij}q_{ij}\sum_i p_{im}q_{ij}} + \frac{\sum_i p_{ij}p_{in}q_{ij}^2}{\sum_i p_{ij}q_{ij}\sum_i p_{in}q_{ij}}\right] + \frac{\sigma_P^2}{4}\left[\frac{\sum_i p_{ij}^2 q_{im}q_{in}}{\sum_i p_{ij}q_{im}\sum_i p_{ij}q_{in}}\right] + \frac{\left(\sigma_P^2 + \sigma_E^2 + \sigma_E^2\sigma_P^2\right)}{4}\left[\frac{\sum_i p_{im}p_{in}q_{ij}^2}{\sum_i p_{im}q_{ij}\sum_i p_{in}q_{ij}}\right]$$

The conditions for structural zeros, etc., are the same as for equations (3) and (4). The derivation of these formulae is given in Appendix 1.

Note that the choice between model 1 and model 2 depends not on which model fits the observed data better, but rather on which model better describes the way the data have been collected. If prices and quantities have been observed directly, then model 1 is the more appropriate. However, if quantities are derived from expenditure data, then the more complex model 2 is relevant.

The formulae (3) to (6) above were applied to the data set used in the 1996 based OECD comparison exercise. (I am grateful to the OECD for making the data available to me.) The full results of the 1996 based OECD comparison exercise are published in OECD (2000).

This data set comprises observations on 207 items, and 32 countries; for the purposes of the present exercise, however, the data on the three balancing items (net changes in stocks, purchases abroad, and net exports), were excluded, so the data set used comprised 204 items. For simplicity, the results have been expressed in terms of correlations, rather than covariances, where the correlation coefficient $\rho(j, m, n)$ is defined as

$$\rho(j,m,n) = \frac{\mathrm{co}(j,m,n)}{\sqrt{v(j,m)v(j,n)}}.$$

Since the values of σ_P^2, σ_Q^2 and σ_E^2 are not known, the formulae were computed using a range of illustrative values of these parameters. Specifically, the implications of "high accuracy" and "low accuracy" variance assumptions were assessed. For the "high accuracy" assumption, the relevant variance parameters were assumed to be 0.000025 (which corresponds to 95 percent of the observations on the given quantity lying within ±1 percent of the true value.) For the "low accuracy" case, the relevant variance parameters were assumed to be 0.0025 (which corresponds to 95 percent of the observations lying within ±10 percent of the true value). All possible combinations of high and low accuracy assumptions were considered; i.e. for model 1, the combinations of (σ_P^2, σ_Q^2) considered were (0.000025, 0.000025), (0.000025, 0.0025), (0.0025, 0.000025) and (0.0025, 0.0025); and correspondingly for (σ_P^2, σ_E^2) under model 2.

For each model, and for each possible pair of variance parameters, the above equations give estimates of 496 variance terms, and 14,880 correlation terms. There is therefore a presentational problem about displaying the results of the calculations in a user-friendly fashion. Fortunately, it turns out that the variance/covariance structure as estimated from the data is relatively simple, and is well described by a small number of summary statistics. These summary statistics are given in Table 1, and are as follows.

Table 1A shows the average log Fisher standard error (that is, the average of the terms $\sqrt{v(j,k)}$), for all possible parameter combinations, and for each model. As would be expected, the average standard errors increase with both variance parameters, for both models. While the standard errors in the two models are fairly comparable when both variance parameters are at the "high accuracy" level, model 2 is rather more sensitive to "low accuracy" in the variance parameters.

Table 1B gives an indication of the spread of the standard error terms, in terms of the ratio of the maximum to the minimum estimated log Fisher standard errors, for each model, and for each pair of parameter values. These ratios indicate that the standard errors are moderately homogeneous: for $\sigma_P^2 = 0.000025$, the

TABLE 1A

AVERAGE LOG FISHER STANDARD ERROR

	Model 1 σ_Q^2		Model 2 σ_E^2	
σ_P^2	0.000025	0.0025	0.000025	0.0025
0.000025	0.0011	0.0105	0.0013	0.0105
0.0025	0.0041	0.0112	0.0081	0.0132

TABLE 1B

RATIO MAX/MIN LOG FISHER STANDARD ERROR

	Model 1 σ_Q^2		Model 2 σ_E^2	
σ_P^2	0.000025	0.0025	0.000025	0.0025
0.000025	1.38	1.3	1.47	1.31
0.0025	2.99	1.4	1.8	1.49

TABLE 1C

AVERAGE LOG FISHER CORRELATION

	Model 1 σ_Q^2		Model 2 σ_E^2	
σ_P^2	0.000025	0.0025	0.000025	0.0025
0.000025	0.45	0.483	0.454	0.483
0.0025	0.244	0.45	0.405	0.453

TABLE 1D

RATIO MAX/MIN LOG FISHER CORRELATION

	Model 1 σ_Q^2		Model 2 σ_E^2	
σ_P^2	0.000025	0.0025	0.000025	0.0025
0.000025	2.18	2.44	1.93	2.43
0.0025	*	2.18	2.32	1.93

*In this case, with high precision quantity measurement, and low precision price measurement under Model 1, there are some negative correlations, so calculation of this ratio is not meaningful.

largest standard error is just less than 50 percent bigger than the smallest; for $\sigma_P^2 = 0.0025$, the largest standard error is usually less than twice the smallest.

All of the estimated correlation terms (excluding structural zeros), were positive, except for the combination of low price accuracy and high quantity accuracy under model 1, when there were a few negative correlations. Table 1C shows the average correlations. Note that these averages are all larger than 0.4,

except for the combination of low price accuracy and high quantity accuracy under model 1. In other words, the correlations between the individual log Fisher terms are indeed typically very material.

Table 1D shows the ratio of the maximum to minimum correlations, except for the case involving the combination of low price accuracy and high quantity accuracy under model 1, for which it is not meaningful to calculate this ratio because there are some negative correlations. It will be noted that the correlations are homogeneous, with the largest typically being about twice the smallest.

Although the detailed results have not been given here, most of the largest standard errors tend to involve one of a relatively small group of countries as one member of the country pair. This group of countries includes Mexico, Russia, Spain, Luxembourg and Hungary.

Overall, the summary statistics illustrate that the variance/covariance structure of the log Fisher indices is relatively simple, under both of the models considered here. The log Fisher standard errors are relatively homogeneous. The correlation terms (other than structural zeros), are typically all positive, and are relatively homogeneous.

The summary statistics also illustrate one feature which has potentially important implications for the strategy of data collection. Under model 2 (that is, when expenditures are measured directly rather than quantities), the log Fisher standard errors are more sensitive to the precision with which prices are collected than under model 1.

The results highlight the need for accuracy in data collection: particularly in collecting price data when model 2 is used. Since one way of increasing sample sizes, and reducing observer error, is through increased use of scanner data for certain items, this suggests that increased collection of scanner data would be useful. In another area, since regional, as opposed to international, price level comparisons are likely to be based on relatively small regional samples, which are therefore likely to be subject to greater imprecision, this points to the need for caution, and to the need to make explicit estimates of precision, in making regional comparisons.

3. Implications for Estimation Techniques

The question then arises: given the estimated variance/covariance structure of the log Fisher indices, what is the optimal GLS estimator corresponding to this structure? There are two reasons for approaching this question cautiously:

(1) The technical difficulty of deriving an exact GLS estimator.

(2) Since knowledge of the variance/covariance structure is based on estimates, it is probably more sensible to look for a robust approximation to the required GLS estimator, which will be close to the optimal estimator, but which will not be subject to the problem of having to recalculate a new GLS estimator from scratch for every new data-set.

The approach adopted here, therefore, is to see whether there is a simplified, "idealized," structure for the variance/covariance matrix, for which it is possible to derive the appropriate GLS estimator; and then to examine what loss of efficiency

is involved in moving from this simplified approximation to the actual variance/covariance structure.

Given that the observed variance/covariance structure, as described in the previous section, has diagonal (variance), terms which are moderately homogeneous, and off-diagonal (correlation), terms that, apart from structural zeros, are positive, and are also usually moderately homogeneous, the natural idealized version of the variance/covariance structure to consider is as follows:

$$\text{var}(y_{jk}) = v^2, \quad \forall j,k, \ \ j \neq k; \tag{7.1}$$

$$\text{corr}(y_{jm}, y_{jn}) = \rho > 0, \forall j,m,n \text{ where } j, \ m \text{ and } n \text{ are all different.} \tag{7.2}$$

$$\text{corr}(y_{jk}, y_{mn}) = 0, \forall j,k,m,n \text{ where } j,k,m \text{ and } n \text{ are all different.} \tag{7.3}$$

The idealized variance/covariance structure described in equations (7) therefore involves constant diagonal (variance) terms, and constant, and positive, off diagonal correlations, apart from the structural zeros implied by (7.3).

The following theorem derives the appropriate GLS estimator, for estimating the parameters in the model at (1) above, when the variance/covariance structure of the error terms in the model is as set out in equations (7).

Theorem 1

Consider the regression model

$$y_{jk} = c_j - c_k + \varepsilon_{jk}, \quad j < k.$$

Define $y_{jj} = 0$ for all j, and $y_{kj} = -y_{jk}$, for $k > j$.
Suppose that the variance/covariance structure of this extended set of $\{y\}$ values satisfies the conditions of equations (7) above.

Then the GLS estimator of $(c_j - c_k)$ is given by

$$\hat{c}_j - \hat{c}_k = \frac{1}{J}\sum_n (y_{jn} - y_{kn}).$$

Proof

The proof of this result, which depends on a symmetry argument, is given in Appendix 2.

The GLS estimator in Theorem 1 is, of course, equivalent to the standard EKS estimator. The importance of this result is that, even though the idealized variance/covariance structure, for $\rho \neq 0$, is very far from satisfying the conditions of uncorrelated errors under which the EKS formula can be derived as noted in para 1.4, nevertheless, the EKS is still optimal for the idealized structure in equations (7).

Note also that, since the argument in the proof of Theorem 1 still holds when $\rho = 0$, Theorem 1 also provides an alternative derivation of the standard EKS formula in the case of uncorrelated errors (though in this case the Theorem 1 proof is much more complicated than is necessary).

It is relatively straightforward, but involves some manipulation, to derive the variance of the estimator $\hat{c}_j - \hat{c}_k$, under the variance assumptions of equations (7). This is given by

$$\text{var}(\hat{c}_j - \hat{c}_k) = \frac{2v^2}{J}[(1-2\rho)+J\rho] \tag{8}$$

This compares with the expression for the variance of the EKS estimator under the assumption of uncorrelated errors, which would be

$$\text{var}(\hat{c}_j - \hat{c}_k) = \frac{2v^2}{J} \tag{9}$$

For $\rho > 0$, the variance at (8) is larger than the variance at (9).

The next question to be considered is: how great is the change in efficiency when the EKS is applied to a model with the variance/covariance structure as estimated in section 2 above, compared with when it is applied with the idealized structure at (7). It is straightforward, but tedious, to show that the variance of the log EKS estimator, under either model 1 or model 2, is given by

$$\text{var}(\log(\text{EKS}_{jk})) = \frac{1}{J^2}\left\{\sum_n [v(j,n)+v(k,n)-2\text{co}(n,j,k)] + \sum_{n_1 \neq n_2} [\text{co}(j,n_1,n_2)+\text{co}(k,n_1,n_2)] + 2v(j,k) + 2\sum_{i \neq j,k} [\text{co}(j,k,i)+\text{co}(k,j,i)]\right\} \tag{10}$$

The ratio of the relative standard errors of the EKS was then calculated, under the estimated and "idealized" structures. For any pair of countries, j and k say, this ratio is given by the (square root of) the ratio of the expressions in formulae (10) and (8) above.

This expression was calculated for each pair of countries, using the values of $v(j, k)$ and $\text{co}(n,j,k)$ as estimated from the data in Section 2, and with values of v^2 and ρ corresponding to the observed average values estimated for the relevant case. A value of this ratio less than 1 means that the EKS is more efficient under the actual variance/covariance structure than under the idealized structure, and vice versa.

Table 2 gives summary statistics for the calculation of these relative efficiency terms. For each case, the lowest and highest values of the above ratio are given. As can be seen, for most of the cases considered, there is little change in the efficiency of the EKS under the actual as compared with the idealized variance/covariance structure. For seven of the eight cases considered, the maximal increase or decrease in efficiency is of the order of 10 percent (and it is worth remarking that what are recorded here are the values of the extreme country pairs; for most of the 496 country pairs considered for each case, the change in efficiency is close to 1). Only for one case, (model 1, $\sigma_P^2 = 0.0025$, $\sigma_Q^2 = 0.000025$), is there a change in efficiency for some country pair exceeding 15 percent.

More detailed examination of the results (not recorded here), shows that there is a fairly consistent pattern, with country pairs involving Russia or Mexico in particular tending to have values of the above ratio consistently above 1 (i.e. the EKS

TABLE 2

RELATIVE EFFICIENCY OF EKS UNDER ACTUAL AND IDEALIZED V/C STRUCTURE

	Model 1 σ_Q^2		Model 2 σ_E^2	
σ_P^2	0.000025	0.0025	0.000025	0.0025
Lowest observed value for any country pair				
0.000025	0.892	0.883	0.925	0.883
0.0025	0.69	0.891	0.923	0.914
Highest observed value for any country pair				
0.000025	1.091	1.092	1.01	1.092
0.0025	1.292	1.089	1.14	1.086

is consistently less efficient for such countries under the actual as opposed to the idealized structure). This corresponds to the fact, as already noted, that the estimated log Fisher variance for country pairs including these countries tends to be relatively large.

Overall, however, the main point to note is that there is generally only a small change in the efficiency of the EKS on moving from the idealized variance/covariance structure to the estimated variance/covariance structure.

This provides strong circumstantial evidence that the EKS estimator is likely to be of close to optimal efficiency for the estimated variance/covariance structure. The argument for this conclusion is as follows. The optimal GLS estimator for the estimated variance/covariance structure is not known, nor is its efficiency. But the efficiency of this estimator is unlikely to be much greater than the efficiency achieved by the EKS estimator for the idealized variance/covariance structure, for which the EKS is known to be optimal. Since the EKS's efficiency does not change much when it is applied to the estimated variance/covariance structure, this implies that the EKS is likely to be close to optimal for this structure too.

The argument above provides a strong case for the application of the EKS estimator to the particular 32 country data set analysed here, compared with any other estimator which is linear in the y_{jk} terms. In addition, another advantage is that application of the EKS does not involve complex calculations to derive an exact GLS estimator each time a new data set is considered; and use of the EKS also means that the estimator used is not sensitive to instability due to the estimation error involved in estimating the variance/covariance structure.

In intuitive terms, these conclusions make a good deal of sense. The group of countries involved in the OECD comparison exercise is a relatively homogeneous group of advanced economies. It is not very surprising that, for a fairly homogeneous group like this, an equally weighted average like the EKS turns out to be close to optimal among all techniques which are linear in the log Fisher indices. Whether similar conclusions apply to more heterogeneous groups of countries would require to be established by further empirical work.

4. Conclusion

The fact that there are errors of estimation inherent in the measurement of prices and quantities (or prices and expenditures), imposes a probabilistic (or stochastic), structure on the set of Fisher indices. Optimal estimation of volume relativities based on the Fisher indices should then take account of this stochastic behavior. If the logarithms of the Fisher indices are considered, then the log Fisher indices will be linear functions of the unknown country volume levels (together with stochastic errors). In this case, the estimation problem is optimally handled by the technique of Generalised Least Squares: the particular GLS estimator which is optimal will be a function of the variance/covariance structure of the log Fisher indices.

The relevance of this paper is as follows:

(1) For two alternative models (relating to two different possible approaches to the collection of the basic data), it derives the relevant formulae for the variance/covariance structures of the individual log Fisher indices.

(2) It applies these formulae to the 1996 OECD data, to give empirical estimates of the variance/covariance structure of the log Fisher indices for this data set; and shows that, for both stochastic models, while there are indeed material correlations between the log Fisher indices, their variance/covariance structure nevertheless has a fairly simple form.

(3) Separately, for an idealized variance/covariance structure which is a fairly close approximation to the variance/covariance structure actually observed, it shows that the standard EKS estimator is in fact the optimal GLS estimator.

(4) It estimates the change in efficiency for the EKS estimator, on moving from the idealized variance/covariance structure (for which the EKS is optimal), to the actual variance/covariance structure; and shows that the change in efficiency is marginal. This provides good circumstantial evidence that the EKS estimator is likely to be close to optimal for this particular data set.

Probably the most striking implication of the above conclusions is the support that these findings give for the EKS method, at least in the context of the specific data set on which the empirical evidence in this paper is based: that is, the OECD96 data set. It should be remembered, however, that these empirical findings relate to this particular data set. Of greater long term relevance are likely to be the theoretical results summarized in (1) and (3) above, which are of general applicability. Also of potential long term significance is the fact that the empirical approach adopted here could be applied to other data sets, to give an indication of how close the EKS is likely to be to optimal efficiency for these data sets.

Appendix A: Derivation of Formulae for Variances and Covariances

(1) The derivation of formulae (3) to (6) makes use of a number of properties of variances and covariances which, for convenience, are listed here first.

(a) If x and y are positive random variables with expected values μ_x and μ_y, then

$$\text{var}(\log(x)) \doteq \frac{\text{var}(x)}{\mu_x^{\,2}}, \quad \text{and} \tag{A1}$$

$$\text{cov}(\log(x), \log(y)) \doteq \frac{\text{cov}(x, y)}{\mu_x \mu_y} \tag{A2}$$

(where "$\doteq$" denotes "is approximately equal to".)

Proof. Expanding $\log(x)$ in a Taylor series implies that

$$\log(x) \doteq \log(\mu_x) + \frac{(x - \mu_x)}{\mu_x},$$

which implies that $E[\log(x)] \doteq \log(\mu_x)$,

and that $[\log(x) - \log(\mu_x)]^2 \doteq \dfrac{(x - \mu_x)^2}{\mu_x^{\,2}}$,

from which (A1) follows on taking expectations.

The proof of (A2) is similar on considering the Taylor expansions of $\log(x)$ and $\log(y)$.

(b) Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be two sets of random variables, where, for $i_1 \neq i_2$, x_{i_1} is independent of x_{i_2}, and y_{i_1} is independent of y_{i_2}. Then

$$\text{var}\left(\sum_i x_i\right) = \sum_i \text{var}(x_i) \tag{A3}$$

$$\text{cov}\left(\sum_i x_i, \sum_i y_i\right) = \sum_i \text{cov}(x_i, y_i) \tag{A4}$$

(These are standard results.)

(c) Let x be a positive random variable with expected value μ_x and variance $\mu_x^{\,2}\sigma^2$: then

$$E(x^{-1}) \doteq \mu_x^{\,-1}, \tag{A5}$$

$$\text{var}(x^{-1}) \doteq \mu_x^{\,-2}\sigma^2. \tag{A6}$$

Proof. Taking a Taylor expansion of x^{-1} implies that

$$\frac{1}{x} \doteq \frac{1}{\mu_x} - \frac{(x - \mu_x)}{\mu_x^{\,2}},$$

from which (A5) follows on taking expectations. Further, it implies that $\left(\dfrac{1}{x} - \dfrac{1}{\mu_x}\right)^2 \doteq \dfrac{(x - \mu_x)^2}{\mu_x^{\,4}}$, from which (A6) follows on taking expectations.

Model 1

(2) Assume the assumptions of model 1 hold. Then

$$
\begin{aligned}
v(j,k) &= \operatorname{var}\big(\log(F_{jk}) = 0.25\operatorname{var}\big(\log\big(F_{jk}{}^2\big)\big)\\
&= 0.25\operatorname{var}\Big[\log\Big(\sum_i p_{ij}q_{ij}\Big)+\log\Big(\sum_i p_{ik}q_{ij}\Big)-\log\Big(\sum_i p_{ij}q_{ik}\Big)-\log\Big(\sum_i p_{ik}q_{ik}\Big)\Big]\\
&= 0.25\Big[\operatorname{var}\Big(\log\Big(\sum_i p_{ij}q_{ij}\Big)\Big)+\operatorname{var}\Big(\log\Big(\sum_i p_{ik}q_{ij}\Big)\Big)+\operatorname{var}\Big(\log\Big(\sum_i p_{ij}q_{ik}\Big)\Big)\\
&\quad+\operatorname{var}\Big(\log\Big(\sum_i p_{ik}q_{ik}\Big)\Big)\Big]+0.5\Big[\operatorname{cov}\Big(\log\Big(\sum_i p_{ij}q_{ij}\Big),\log\Big(\sum_i p_{ik}q_{ij}\Big)\Big)\\
&\quad-\operatorname{cov}\Big(\log\Big(\sum_i p_{ij}q_{ij}\Big),\log\Big(\sum_i p_{ij}q_{ik}\Big)\Big)-\operatorname{cov}\Big(\log\Big(\sum_i p_{ij}q_{ij}\Big),\log\Big(\sum_i p_{ik}q_{ik}\Big)\Big)\\
&\quad-\operatorname{cov}\Big(\log\Big(\sum_i p_{ik}q_{ij}\Big),\log\Big(\sum_i p_{ij}q_{ik}\Big)\Big)-\operatorname{cov}\Big(\log\Big(\sum_i p_{ik}q_{ij}\Big),\log\Big(\sum_i p_{ik}q_{ik}\Big)\Big)\\
&\quad+\operatorname{cov}\Big(\log\Big(\sum_i p_{ij}q_{ik}\Big),\log\Big(\sum_i p_{ik}q_{ik}\Big)\Big)\Big]
\end{aligned}
$$

The evaluation of a typical term in this expression is illustrated: for example,

$$
\operatorname{var}\Big(\log\Big(\sum_i p_{ij}q_{ij}\Big)\Big)\doteq\frac{\operatorname{var}\Big(\sum_i p_{ij}q_{ij}\Big)}{\Big(\mathrm{E}\Big(\sum_i p_{ij}q_{ij}\Big)\Big)^2},\ \text{by (A1).}
$$

Now $\operatorname{var}\Big(\sum_i p_{ij}q_{ij}\Big)=\sum_i \operatorname{var}(p_{ij}q_{ij})$, by (A3):

also $E(p_{ij}q_{ij})=E(\dot p_{ij}\dot q_{ij}(1+\phi_{ij})(1+\gamma_{ij}))=\dot p_{ij}\dot q_{ij}$,

and $E\big[(p_{ij}q_{ij})^2\big]=E\big[\dot p_{ij}{}^2\dot q_{ij}{}^2(1+\phi_{ij})^2(1+\gamma_{ij})^2\big]$

$=\dot p_{ij}{}^2\dot q_{ij}{}^2E\big[(1+\phi_{ij}{}^2+\gamma_{ij}{}^2+\phi_{ij}{}^2\gamma_{ij}{}^2+\text{terms of first order in }\phi\text{ or }\gamma\big]$

$=\dot p_{ij}{}^2\dot q_{ij}{}^2\big(1+\sigma_P{}^2+\sigma_Q{}^2+\sigma_P{}^2\sigma_Q{}^2\big)$,

so $\operatorname{var}(p_{ij}q_{ij})=E\big[(p_{ij}q_{ij})^2\big]-[E(p_{ij}q_{ij})]^2$

$=\dot p_{ij}{}^2\dot q_{ij}{}^2\big(\sigma_P{}^2+\sigma_Q{}^2+\sigma_P{}^2\sigma_Q{}^2\big)$.

It follows that

$$
\operatorname{var}\Big(\log\Big(\sum_i p_{ij}q_{ij}\Big)\Big)\doteq\frac{\big(\sigma_P{}^2+\sigma_Q{}^2+\sigma_P{}^2\sigma_Q{}^2\big)\sum_i \dot p_{ij}{}^2\dot q_{ij}{}^2}{\Big[\sum_i \dot p_{ij}\dot q_{ij}\Big]^2}:
$$

however, the terms $\dot p_{ij}$ and $\dot q_{ij}$ are unknown: so as a final approximation, $\dot p_{ij}$ and $\dot q_{ij}$ are replaced by p_{ij} and q_{ij} in this expression, giving

$$\operatorname{var}\left(\log\left(\sum_i p_{ij}q_{ij}\right)\right) \doteq \frac{(\sigma_P{}^2 + \sigma_Q{}^2 + \sigma_P{}^2\sigma_Q{}^2)\sum_i p_{ij}{}^2 q_{ij}{}^2}{\left[\sum_i p_{ij}q_{ij}\right]^2}.$$

Each of the terms in the above expression for $v(j, k)$ can be evaluated similarly, (using (A2) and (A4) at the appropriate stages for the covariance terms, rather than (A1) and (A3)), hence establishing formula (3).

(3) The derivation of formula (4) for the covariance $\mathrm{co}(j, m, n)$ proceeds by an exactly analogous argument.

Model 2

(4) The derivation of $v(j, k)$ for model 2 starts with the same expression for $v(j, k)$ as set out at the beginning of paragraph (2) above. The derivation of a typical term in this expression is illustrated, under the assumptions of model 2, as follows. The term chosen for illustrative purpose is $\operatorname{var}\left(\log\left(\sum_i p_{ij}q_{ik}\right)\right)$. As before,

$$\operatorname{var}\left(\log\left(\sum_i p_{ij}q_{ik}\right)\right) \doteq \frac{\operatorname{var}\left(\sum_i p_{ij}q_{ik}\right)}{\left(E\left(\sum_i p_{ij}q_{ik}\right)\right)^2}, \quad \text{by (A1)},$$

$$\text{and } \operatorname{var}\left(\sum_i p_{ij}q_{ik}\right) = \sum_i \operatorname{var}(p_{ij}q_{ik}), \quad \text{by (A3)};$$

so what is required is to evaluate $E(p_{ij}q_{ik})$ and $\operatorname{var}(p_{ij}q_{ik})$ under the assumptions of model 2.

Now, under model 2, q_{ij} is equal to $e_{ij}p_{ij}{}^{-1}$; moreover, by (A5) and (A6), $p_{ij}{}^{-1} \doteq \dot{p}_{ij}{}^{-1}(1 + \mu_{ij})$, say, where the error term μ_{ij} has variance $\sigma_P{}^2$.

Hence

$$\begin{aligned} E(p_{ij}q_{ik}) &= E\left(p_{ij}e_{ik}p_{ik}{}^{-1}\right) \doteq E\left(\dot{p}_{ij}\dot{e}_{ik}\dot{p}_{ik}{}^{-1}(1+\phi_{ij})(1+\lambda_{ik})(1+\mu_{ik})\right) \\ &= \dot{p}_{ij}\dot{e}_{ik}\dot{p}_{ik}{}^{-1} = \dot{p}_{ij}\dot{q}_{ik}, \end{aligned}$$

and

$$\begin{aligned} E\left[(p_{ij}q_{ik})^2\right] &= E\left(p_{ij}{}^2 e_{ik}{}^2 p_{ik}{}^{-2}\right) \doteq E\left(\dot{p}_{ij}{}^2\dot{e}_{ik}{}^2\dot{p}_{ik}{}^{-2}(1+\phi_{ij})^2(1+\lambda_{ik})^2(1+\mu_{ik})^2\right) \\ &= \dot{p}_{ij}{}^2\dot{e}_{ik}{}^2\dot{p}_{ik}{}^{-2}E\left[(1+\phi_{ij}{}^2)(1+\lambda_{ik}{}^2)(1+\mu_{ik}{}^2) + \text{terms of first order in } \phi, \lambda \text{ or } \mu\right] \\ &= \dot{p}_{ij}{}^2\dot{e}_{ik}{}^2\dot{p}_{ik}{}^{-2}\left(1+\sigma_P{}^2\right)^2\left(1+\sigma_E{}^2\right) \\ &= \dot{p}_{ij}{}^2\dot{q}_{ik}{}^2\left(1+\sigma_P{}^2\right)^2\left(1+\sigma_E{}^2\right). \end{aligned}$$

Hence $\operatorname{var}(p_{ij}q_{ik}) \doteq \left[\left(1+\sigma_P{}^2\right)^2\left(1+\sigma_E{}^2\right) - 1\right]\dot{p}_{ij}{}^2\dot{q}_{ik}{}^2$,

and thus

$$\operatorname{var}\left(\log\left(\sum_i p_{ij}q_{ik}\right)\right) \doteq \frac{\left[\left(1+\sigma_P{}^2\right)^2\left(1+\sigma_E{}^2\right)-1\right]\sum_i p_{ij}{}^2 q_{ik}{}^2}{\left[\sum_i p_{ij}q_{ik}\right]^2}.$$

The other terms in $v(j, k)$ are evaluated similarly, so establishing formula (5): the derivation of formula (6) for the covariance $\mathrm{co}(j, m, n)$ proceeds by an exactly analogous argument.

Appendix B: Proof of Theorem 1

(1) Consider the regression model

$$y_{jk} = c_j - c_k + \varepsilon_{jk}, \quad j < k.$$

The parameters in this model are not identified, up to an additive constant. It is sufficient to identify the parameters to impose the single identifiability constraint $\sum_j c_j = 0$. Subject to this constraint, the model is then fully identified; so, by the theory of GLS, there exists a unique minimum variance linear unbiased estimator of the parameters. This means, in particular, that the estimator of c_1, denoted as $\hat{c}_1$, can be written in the form

$$\hat{c}_1 = \sum_{k>j} a_{jk} y_{jk},$$

for some constants a_{jk}, where the terms a_{jk} are uniquely defined (and do not depend on the y values).

(2) Now, consider what happens if, keeping the position of country 1 fixed, the order in which the remaining countries has been numbered is permuted. (For example, instead of writing, say, USA = country number 1, France = country number 2, Germany = country number 3, . . . , instead USA = country number 1, Germany = country number 2, = France country number 3.) Rearranging the ordering of countries, like this, gives an alternative expression of the original regression model, with a new parameter vector, c^*, say, which is a permutation of the original parameter vector c (and with $c_1{}^* = c_1$, since the position of country 1 has not been altered); with a new observation vector, y^* (which is a function of the original observation vector y); and with a new error vector, ε^* (which is a function of the original error vector ε).

(3) The special feature which is the key to the proof of Theorem 1 is that, in terms of the new ordering of the countries, this new regression model can be written

$$y_{ji}{}^* = c_j{}^* - c_k{}^* + \varepsilon_{jk}{}^*, \quad j < k.$$

In other words, the new model can be written as

$$y^* = Xc^* + \varepsilon^*,$$

where the design matrix, X, is identical to the design matrix of the original regression model $y = Xc + \varepsilon$. Moreover, it is clear from the equations (7) that the variance/covariance matrices of the vectors ε and ε^* are identical; and, further, since the c^* terms are a permutation of the c terms, the c^* terms satisfy the constraint

$$\sum_j c_j^* = 0.$$

In other words, the re-written regression model is formally identical (having the same design matrix, variance/covariance structure and identifiability constraint), to the original model: so, in particular, it follows that

$$\hat{c}_1^* = \sum_{k>j} a_{jk} y_{jk}^*,$$

where the coefficients a_{jk} are the same as the coefficients in the definition of $\hat{c}_1$ in paragraph (1) above.

Moreover, permuting the ordering of later countries will clearly have no effect on the estimation of the parameter relating to country 1: in other words, $\hat{c}_1^* = \hat{c}_1$. That is, the following identity must hold, namely

(B1)
$$\sum_{k>j} a_{jk} y_{jk}^* = \sum_{k>j} a_{jk} y_{jk},$$

where identity (B1) must hold for all possible observation vectors y (and the terms a_{jk} do not depend on y).

(4) Now suppose that the particular permutation involved in the alternative regression model has involved switching the positions of countries m and $(m + 1)$, for some $m > 1$, leaving the positions of all other countries in the ordering fixed. Then, for this particular permutation, the terms y_{jk}^* are defined in terms of the terms y_{jk} by the following relationships:

$$\begin{aligned}
&y_{jk}^* = y_{jk}, \quad \forall j,k \neq m \text{ or } (m+1),\\
&y_{jm}^* = y_{j(m+1)}, \quad \forall j < m,\\
&y_{m(m+1)}^* = -y_{m(m+1)},\\
&y_{mj}^* = y_{(m+1)j}, \quad \forall j > m+1,\\
&y_{j(m+1)}^* = y_{jm}, \quad \forall j < m,\\
&y_{(m+1)j}^* = y_{mj}, \quad \forall j > m+1.
\end{aligned}$$

(5) Recall that the identity at (B1) holds for all possible observation vectors y. Consider the hypothetical observation vector y which consists entirely of zeroes, apart from the single term $y_{m(m+1)}$, which is 1. Substituting this y vector (and the corresponding values of y^* from paragraph (4), into (B1), it follows that

$$-a_{m(m+1)} = a_{m(m+1)},$$

that is,

(B2) $$a_{m(m+1)} = 0.$$

(6) For a selected $j < m$, consider the hypothetical observation vector y which consists entirely of zeroes, apart from the single term $y_{j(m+1)}$, which is 1. Substituting this y vector (and the corresponding values of y^* from paragraph (4), into (B1), it follows that

(B3) $$a_{jm} = a_{j(m+1)} \quad \forall j < m.$$

(7) Since m was selected arbitrarily in the range $1 < m < J$, it follows that (B2) and (B3) hold for all m in this range. (B3) implies that the upper triangular array $\{a_{jk}, k > j\}$, is constant along its rows: and (B2) implies that all the rows of this array, apart from the first row, must consist of zeros. In other words, $\hat{c}_1$ is of the form

$$\hat{c}_1 = a\sum_{k>1} y_{1k}.$$

(8) Another permutation argument is now used to derive the general form of $\hat{c}_j$. Suppose that countries (1) and (j) are permuted in the original selection of countries, giving a new set of observations denoted as y^*, and a new parameter vector c^* (where $c_1^* = c_j$). Again, this permutation has no effect on the design matrix, covariance structure, or identifiability constraint of the regression model, so it follows that

$$\hat{c}_j = \hat{c}_1^* = a\sum_{k>1} y_{1k}^*.$$

However,

$$\begin{aligned} y_{1k}^* &= -y_{kj} \quad \text{for} \quad 1 < k < j, \\ y_{1j}^* &= -y_{1j}, \\ y_{1k}^* &= y_{jk} \quad \text{for} \quad k > j. \end{aligned}$$

So

$$\hat{c}_j = \hat{c}_1^* = a\sum_{k>1} y_{1k}^* = a\left[-\sum_{k<j} y_{kj} + \sum_{k>j} y_{jk}\right], \quad \text{for all } j.$$

Taking the expectation of this expression, it follows that

$$\begin{aligned} E(\hat{c}_j) &= a\left[(J-1)c_j - \sum_{k\neq j} c_k\right] \\ &= a\left[Jc_j - \sum_k c_k\right]. \end{aligned}$$

Therefore $E(\hat{c}_j - \hat{c}_k) = aJ[c_j - c_k]$:
so $(\hat{c}_j - \hat{c}_k)$ is an unbiased estimator of $(c_j - c_k)$, (which it must be), if, and only if $a = \dfrac{1}{J}$. This establishes the result.

REFERENCES

Balk, Bert M., "Axiomatic Price Theory: A Survey," *International Statistical Review*, 63, 1, 69–93, 1995.

———, "Aggregation Methods in International Comparisons, What Have We Learned," *Erasmus Institute of Management Report Series*, June 2001.

Cuthbert, James R., and Margaret Cuthbert, "On Aggregation Methods of Purchasing Power Parities," *OECD Department of Economics and Statistics Working Papers*, No. 56, 1988.

Diewert, W. Erwin, "On the Stochastic Approach to Index Numbers," *University of British Columbia Department of Economics Discussion Paper*, 95–31, 1995.

Elteto, O. and P. Koves, "On an Index Computation Problem in International Comparisons," *Statiztikai Szemle*, 42, 507–18, 1964.

Gini, C., "Quelques Considerations au Sujet de la Construction des Nombres Indices des Prix et des Questions Analogues," *Metron*, 4, 3–162, 1924.

Hill, Robert J., "A Taxonomy of Multilateral Methods for Making International Comparisons of Prices and Quantities," *Review of Income and Wealth*, 43, 49–69, 1997.

Johnston, Jack and John Dinardo, *Econometric Methods, Fourth Edition*, McGraw-Hill, New York, 1997.

OECD (Organization for Economic Co-operation and Development), *Purchasing Power Parities and Real Expenditures, 1996 Results*, Paris, 2000.

Rao, Prasada, "Weighted EKS Generalised CPD Method for Aggregation at Basic Heading Level and above Basic Heading Level," *Joint World Bank/OECD Seminar on Recent Advances in Purchasing Power Parities*, Washington, January/February 2001.

Szulc, Bodan, "Index Numbers of Multilateral Regional Comparisons," *Przeglad Statystyczny*, 11, 239–54, 1964.