Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

ESTIMATING SAMPLING ERRORS FOR MOVEMENTS IN THE UK INDEX OF PRODUCTION

Susan Full, Daniel Lewis[1]

ABSTRACT

Methodology for estimating the sampling error of the non-seasonally adjusted estimate of level of the Index of Production (IoP) has previously been developed using Taylor linearisation and parametric bootstrap methods with both producing comparable results. From the study it was considered that the parametric bootstrap would be more practical to implement. This paper describes the methodology that is being developed to estimate the sampling error of the non-seasonally adjusted IoP change using the parametric bootstrap, along with the data that is needed from the contributing surveys, the assumptions made and the practical problems met during development.

KEY WORDS: sampling error; Index of Production; parametric bootstrap.

1. INTRODUCTION

1.1 The Index of Production

The Index of Production (IoP) is one of the main economic indicators produced by the UK Office for National Statistics (ONS). It is a monthly index of the output of the production industries and is published in its own right as well as being a component of the output measure of GDP, GDP(O). The index is constructed using data from several different sources, most are ONS surveys but there are some external sources. The main ONS surveys are the Monthly Production Inquiry (MPI), Producer Price Index (PPI), Export Price Index (EPI) and the Quarterly Stocks Inquiry (QSI). Externally sourced data include volume data for oil, gas, electricity and mining industries.

1.2 Sampling errors for the IoP

Currently sampling errors (SEs) are not calculated for the IoP although the need for them has long been recognised. Previous work developed methods to estimate sampling errors for the level of the non-seasonally adjusted IoP and is fully described in Kokic (1998). This work approached the problem using both Taylor linearisation and a parametric bootstrap technique (Efron and Tibshirani, 1993). Both methods produced comparable results. The recommendation was to use the parametric bootstrap for practical reasons; it is a more flexible method and avoids the need for complex mathematical derivations.

[1] Susan Full, Daniel Lewis, Office for National Statistics, Government Buildings, Cardiff Road, NEWPORT, UK, NP10 8XG
Acknowledgements: The authors acknowledge the work of Phil Kokic in developing the methodology for this work.

However, the estimates of movement in the IoP are considered to be more important than estimates of level. Fortunately the technique in Kokic (1998) is sufficiently general to allow it to be used to estimate SEs of movement, as well as other more complicated statistics derived from the IoP time series. Therefore, further work has been carried out under contract to ONS to develop the bootstrap technique to produce SEs for the movement as well as the level.

The methodology proposed was developed in Kokic (2001), and then ONS applied it to IoP data.

2. METHODOLOGY

2.1 Definitions

The inputs to the IoP compilation are defined as:

$\hat{D}_{0t,1h}$	is the PPI home price deflator between times 0 and *t*,
$\hat{D}_{0t,2h}$	is the export price deflator,
$\hat{D}_{0th}$	is the combined deflator,
$\hat{S}_{th}, \hat{S}_{t1h}, \hat{S}_{t2h}$	are the total sales, home sales and export sales for month *t* estimated from the MPI,
$\hat{U}_{th}$	is the change in value of stocks for one quarter (closing minus opening levels of work-in-progress plus finished goods) estimated from the QSI.
$\hat{g}_{0h}$	is the group divisor estimated from the MPI as the monthly average of total sales in the base year, and
w_{0h}	gross value added in the base year estimated in the annual business inquiry (ABI).

Where subscript *h*, represents the 4-digit Standard Industrial Classification (SIC). Also, note that the IoP is produced monthly, and so the subscript *t* represents month. For the quarterly data from QSI, the change in stocks from the quarter that *t* falls in is used. For the base time point, *t=0*, an annual average is used.

It was shown in Kokic (2000) that a good approximation to the (non-seasonally adjusted) IoP at the 4-digit SIC level, at least appropriate for estimating its sampling variance, is (2.1). The approximation excludes a number of terms in order to simplify the equation.

$$I_{0th} = \frac{1}{\hat{g}_{0h}}\left(\frac{\hat{S}_{th}}{\hat{D}_{0th}} + \frac{\hat{U}_{th}}{3\hat{D}_{0t,1h}}\right), \tag{2.1}$$

where

$$\hat{D}_{0th} = \hat{S}_{th}\left(\frac{\hat{S}_{t1h}}{\hat{D}_{0t,1h}} + \frac{\hat{S}_{t2h}}{\hat{D}_{0t,2h}}\right)^{-1}. \tag{2.2}$$

It follows that

$$I_{0th} = \frac{1}{\hat{g}_{0h}}\left(\frac{\hat{S}_{t1h}}{\hat{D}_{0t,1h}} + \frac{\hat{S}_{t2h}}{\hat{D}_{0t,2h}} + \frac{\hat{U}_{th}}{3\hat{D}_{0t,1h}}\right). \tag{2.3}$$

For higher levels, the index is computed as a weighted average over the appropriate industries according to gross value added in the base year:

$$I_{0t} = \frac{\sum_h w_{0h} I_{0th}}{\sum_h w_{0h}} . \tag{2.4}$$

The relative change in the IoP between two time points, $0 < r < t$ say, is computed as:

$$I_{rt} = \frac{I_{0t}}{I_{0r}} . \tag{2.5}$$

2.2 Methodology for estimating the SEs

This section describes how to use the parametric bootstrap to estimate the sampling errors of I_{rt} .

Let

$$\hat{\mu}_{rth} = \left(\hat{S}_{r1h}, \hat{S}_{r2h}, \hat{S}_{t1h}, \hat{S}_{t2h}, \hat{g}_{0h}, \hat{D}_{0r,1h}, \hat{D}_{0r,2h}, \hat{D}_{0t,1h}, \hat{D}_{0t,2h}, \hat{U}_{rh}, \hat{U}_{th}\right)' \tag{2.6}$$

and $\hat{\Sigma}_{rth}$ be the estimated variance-covariance matrix corresponding to $\hat{\mu}_{rth}$. Thus the diagonal elements of $\hat{\Sigma}_{rth}$ are the variance estimates of the corresponding elements of $\hat{\mu}_{rth}$, and the off-diagonal elements are the covariance estimates. In this section we assume that all these estimates are available. However, it is not necessary to estimate all covariances. As explained in the next section many of these can be set to zero without any substantial risk of introducing bias.

From (2.3) we see that the 4-digit SIC IoP index can be represented as a function of $\hat{\mu}_{rth}$:

$$\begin{aligned} I_{0th} &= I_{0th}(\hat{\mu}_{rth}) = \frac{1}{\hat{g}_{0h}}\left(\frac{\hat{S}_{t1h}}{\hat{D}_{0t,1h}} + \frac{\hat{S}_{t2h}}{\hat{D}_{0t,2h}} + \frac{\hat{U}_{th}}{3\hat{D}_{0t,1h}}\right) \\ I_{0rh} &= I_{0rh}(\hat{\mu}_{rth}) = \frac{1}{\hat{g}_{0h}}\left(\frac{\hat{S}_{r1h}}{\hat{D}_{0r,1h}} + \frac{\hat{S}_{r2h}}{\hat{D}_{0r,2h}} + \frac{\hat{U}_{rh}}{3\hat{D}_{0r,1h}}\right). \end{aligned} \tag{2.7}$$

Consequently, the bootstrap algorithm for estimating the SE of I_{rt} is as follows.

a) For, $b=1,\ldots,B=200$ say, perform steps (b)-(e).

b) For each h, randomly sample a value from a multivariate normal distribution with mean $\hat{\mu}_{rth}$ and covariance $\hat{\Sigma}_{rth}$. Denote the simulated values by $\hat{\mu}_{rth(b)}$.

c) Using (2.7) compute $I_{0rh(b)} = I_{0rh}\left(\hat{\mu}_{rth(b)}\right)$ and $I_{0th(b)} = I_{0th}\left(\hat{\mu}_{rth(b)}\right)$.

d) Weighting as in (2.4), determine $I_{0r(b)}$ and $I_{0t(b)}$.

e) Compute $I_{rt(b)} = \frac{I_{0t(b)}}{I_{0r(b)}}$.

f) Estimate the SE of I_{rt} using the bootstrap variance formula:

$$SE\left(I_{rt}\right)^2 \equiv \hat{v}\left(I_{rt}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(I_{rt(b)} - \bar{I}_{rt}\right)^2,$$

where $\bar{I}_{rt} = B^{-1}\sum_{b} I_{rt(b)}$.

In step (b) the use of a multivariate normal distribution to simulate the bootstrap values of $\hat{\mu}_{rth}$ is based on the fact that, according to the central limit theorem, the distribution of $\hat{\mu}_{rth}$ is close to multivariate normal. The robustness of this assumption was tested by simulation in Kokic (1998) and the results were generally good.

2.3 Structure of the covariance matrix

As mentioned in the previous section, the matrix $\hat{\Sigma}_{rth}$ has some structure which reduces significantly the number of covariances that need to be estimated. Furthermore, following the approach of Kokic (1998), one is able to make some simplifying assumptions which will further reduce the number of inputs required. These assumptions will also be outlined in the current section.

ONS uses synchronised PRN sampling to select samples for its business surveys. The period of rotation is currently set to 15 months for monthly surveys (e.g. MPI), and 5 quarters for quarterly surveys (e.g. QSI). The PPI has a more complicated rotation system because it is a two-phase design; see Hedlin (2000) for more details. From this we make the following assumptions:

- sample selection processes for each survey are (approximately) statistically independent of each other;
- estimates that are more than 15 months apart are assumed to be uncorrelated.

Given these assumptions, provided *r* and *t* are at least 15 months apart, it can be assumed that $\hat{\Sigma}_{rth}$ is diagonal with no correlations (except for the deflators from the PPI).
More generally, for any time difference one can assume the following structure for the covariance matrix:

$$\hat{\Sigma}_{rth} = \begin{pmatrix} \Sigma_{rth}^{(1)} & 0 & 0 & 0 & 0 \\ 0 & \Sigma_{rth}^{(2)} & 0 & 0 & 0 \\ 0 & 0 & \Sigma_{rth}^{(3)} & 0 & 0 \\ 0 & 0 & 0 & \Sigma_{rth}^{(4)} & 0 \\ 0 & 0 & 0 & 0 & \Sigma_{rth}^{(5)} \end{pmatrix}, \tag{2.8}$$

where

$\Sigma_{rth}^{(1)}$ is a 4×4 covariance matrix for $\left(\hat{S}_{r1h}, \hat{S}_{r2h}, \hat{S}_{t1h}, \hat{S}_{t2h}\right)'$,
$\Sigma_{rth}^{(2)}$ is a 1×1 covariance matrix for $\hat{g}_{0h}$ (the variance of $\hat{g}_{0h}$),
$\Sigma_{rth}^{(3)}$ is a 2×2 covariance matrix for $\left(\hat{D}_{0r,1h}, \hat{D}_{0t,1h}\right)'$,
$\Sigma_{rth}^{(4)}$ is a 2×2 covariance matrix for $\left(\hat{D}_{0r,2h}, \hat{D}_{0t,2h}\right)'$, and
$\Sigma_{rth}^{(5)}$ is a 2×2 covariance matrix for $\left(\hat{U}_{rh}, \hat{U}_{th}\right)'$.

This form is appropriate for use in practice, but it still requires the estimation of 11 variances plus 9 covariances for each category *h*. Implicitly, estimating the variance of the IoP using the approach outlined in section 2.2 and this covariance matrix involves two basic assumptions:

- The gross value-added weights, w_{0h}, are fixed and do not have any sampling error.
- Both *r* and *t* are at least 12 months after the base year so that there is no correlation between the group divisor and the other home sales estimates.

The first of these would usually be considered standard in index construction, although it could be worked around by adding the weight as an extra element to $\hat{\mu}_{rth}$. The second assumption is somewhat restrictive, particularly if a chain-linking is used for the IoP in the future. In addition, Kokic (1998) made the assumption that:

- $\text{relative variance}\left(\hat{g}_{0h}\right) = \dfrac{\text{relative variance}\left(\hat{S}_{th}\right)}{4}$.

The impact of a false value for the divisor on the right hand side of this relationship was found to have very little impact on the SE of the IoP. Another assumption made by Kokic (1998) was that for the variance estimates, $\hat{v}$, of the home price deflator and the export price deflator, $\hat{v}\left(\hat{D}_{0t,2h}\right) = \hat{v}\left(\hat{D}_{0t,1h}\right)$. In fact the export price deflator does not have a sampling variance. It was argued in Kokic (1998) that such an assumption should, to a reasonable extent, account for possible bias in the IoP resulting from the export price deflator. The corresponding assumption in this more general situation is:

- $\Sigma_{rth}^{(4)} = \Sigma_{rth}^{(3)}$.

Finally, since the same stock adjustment is used in any given quarter, in the case where *r* and *t* are from the same quarter, $\hat{U}_{rh} = \hat{U}_{th}$. In other words, the number of elements in the $\hat{\mu}_{rth}$ vector can sometimes be reduced by one and the $\hat{\Sigma}_{rth}$ matrix can be slightly simplified.

3. APPLICATION

3.1 Outline

The main emphasis of the project within the ONS has been to take the proposed methodology and to apply it to survey data in order to test the applicability of the methodology and to produce some estimates of SEs. Initially, only a sub-set of industries was included in the development but it is now being extended to the whole production economy, for those industries that do not rely on externally sourced data.

3.2 Data requirements

For each component of the IoP calculation the estimate of the level for each period, variances and covariances are needed. For most ONS business surveys variances are routinely calculated but covariances have not been available. A separate sub-project has considered sampling errors for movements in the business surveys, with sample rotation and a dynamic population, and in order to do this it was necessary to estimate the covariance. This work is described in Full and Lewis (2001).

Some of the constituent surveys are not stratified in the same way as the IoP industry stratification. For instance, the PPI estimates are available at a 6-digit product code level and for each IoP 4-digit SIC a number of PPIs are needed. Also, for the QSI the data is collected at a higher level of aggregation and therefore the stocks estimates have to be split appropriately. For these components suitable weights were also needed in order to produce the necessary estimates. An additional problem with the QSI is that the survey uses a non-standard estimation system, and therefore it has been difficult to produce appropriate variance and covariance estimates for this component. Results so far have either ignored the input of the QSI, or made assumptions about the variances and covariances.

3.3 Simulating observations from a multivariate normal distribution

One of the crucial steps in the algorithm presented in section 2.2 is step (b), the random sampling of an observation from a multivariate normal distribution. This section describes the method suggested in Kokic (2001).

Most statistical packages do not include a routine for directly performing this sampling. However, they usually have routines for pseudo-random sampling from a univariate standard normal random variable. A single *k*-dimensional multivariate normal random variable with mean μ and covariance matrix Σ can be constructed by an appropriate transformation of *k* independent univariate standard normal values, $Z_1,\ldots,Z_k$, say. The same transformation can be used to produce a random sample from the multivariate normal random variable.

The key to performing this task is the construction of a linear transformation of the Z's with the appropriate mean and covariance. Note that $Z=(Z_1,\ldots,Z_k)'$ is a multivariate normal random variable with mean 0 and covariance *I* (the $k\times k$ identity matrix). If *A* is a $k\times k$ matrix such that $AA'=\Sigma$, then it is easy to show that

$$Y=\mu+AZ$$

has a multivariate normal distribution with mean μ and covariance Σ. The *Cholesky decomposition* is one method that can be used for constructing *A*, see Press et al. (1992).

3.4 Application of the method

To use this method Σ needs to be symmetric and positive definite. These are inherent properties of the variance-covariance matrices of random vectors, see Harville (1997). However, because we are estimating the components of the variance-covariance matrices, there are inconsistencies between variances and covariances for some industries. This has led to difficulties in meeting the condition of positive definiteness. A method[2] has been developed which gives a solution to the problems of invalid correlations and first order partial correlations in a variance-covariance matrix, that is correlations and first order partial correlations with an absolute value greater than one. The technique alters covariance values until the matrix is consistent to this level. For more details see Wood (2001). This technique solves many of the problems, but some nonpositive definite variance-covariance matrices remain. These arise because of the small sample sizes used, in particular, in the MPI variables. In some industries this means that it is impossible to calculate valid variances and covariances. In the short term we have solved this by imputing variances and covariances from related industries.

[2] The authors would like to acknowledge the contribution of Markus Šova of ONS to this work.

3.5 Number of bootstrap replications

The proposed methodology suggested that 200 replications of the bootstrap should be sufficient to estimate the SE. During development we have found that in many cases this has not been sufficient. Therefore, we have considered introducing a stopping rule to ensure that enough replications are used. We have also found that plotting estimates of the SE, figure 1, against number of replications has been a useful debugging tool. Plots produced for the data used so far suggest that 1000 replications are necessary to give convergence to within ± 0.5 % of the coefficient of variation.

Plots of this kind have also been very useful in highlighting problems with the quality of the data.

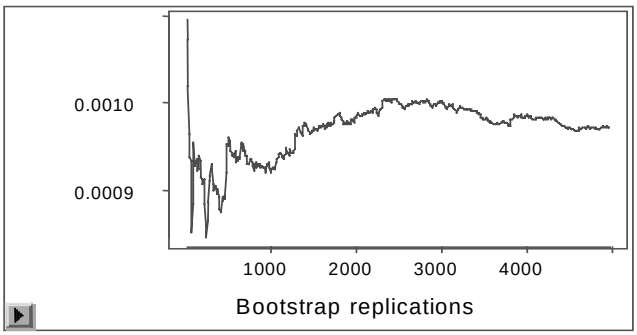

Figure 1 Estimate of variance of change for one industry plotted against number of replications.

4. FURTHER WORK

4.1 A chain-linked version of the IoP

Chain-linking of the IoP will be introduced in the next few years and consideration has been given to the applicability of the proposed method. The conclusions in Kokic (2001) were that:

- The crucial step for the inclusion of a new index in the current methodology is the specification of the function (2.3) linking $\hat{\mu}_h$ to the chain-linked IoP. Depending on which level linking will be performed at, it may be necessary to redefine for a level other than 4-digit SIC; see Tuke and Reed (2001).
- In re-specifying (2.3) it is not necessary to be able to write the exact mathematical formula for the index; an algorithmic specification is sufficient. In fact, an algorithmic specification may be simpler for the chain-linked index.
- Regardless of the way the new index is defined, it will still only be a function of the survey estimates in $\hat{\mu}_h$. For this reason the general algorithm presented in Kokic (2001) (with appropriate modification), will continue to be valid.

4.2 Other issues

Other issues that still need to be considered include how the methodology should be implemented in order to produce SEs to meet user requirements within operating constraints. In particular, how often SEs should be estimated and also whether the IoP SE estimates should be smoothed or whether smoothing should be applied to the variances and covariances of the constituent surveys. It has also been suggested that the use of General Variance Functions might be suitable for this application. The aim of the current project is to produce SEs for the monthly movement but users are also interested in the movement of three-months over the previous three-months, and three-months on three-months a year previous. Using the

methodology proposed it should be possible to also estimate SEs for these statistics. One issue that might need to be addressed will be whether the enlarged variance-covariance matrix will cause additional problems in meeting the conditions necessary for applying the Cholesky decomposition.

REFERENCES

Berger, Y. (2000). "Multilevel estimation of generalised variance functions". Paper presented at the 15th Australian Statistical Conference, Adelaide, Australia. July, 2000.

Efron, B. and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman and Hall, New York, London.

Full, S. and Lewis, D. (2001) "Estimating Sampling Errors for Movements in Business Surveys", Proceedings of the Quality in Official Statistics conference, Statistics Sweden.

Harville, D. A. (1997). Matrix Algebra From A Statistician's Perspective. Springer-Verlag, New York.

Hedlin, D. (2000). "A discussion of the estimation for the Producer Price Index". Internal ONS report.

Kokic, P. N. (1998). "Estimating the sampling variance of the UK Index of Production". Journal of Official Statistics, 14, 163-179.

Kokic, P. (2000). "Incorporation of the Stock Adjustment in the Sampling Variance of the IoP". National Statistics Quality Review Series No. 1: Review of Short-Term Output Indicators, ONS.

Kokic, P. N. (2001) "Estimating the Sampling Variance of a Change in the IoP", unpublished report, Office for National Statistics.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). Numerical Recipes in C: The Art of Scientific Computing, Second edition. Cambridge University Press.

Purdon, S., Cope, I. and Davies, G. (1998). "Obtaining an optimal allocation and weighting system for the United Kingdom Producer Price Indices". Internal ONS report.

Tuke, A. and Reed, G. (2001). "The effects of annual chain-linking on the output measure of GDP". Economic Trends, October.

Wood, J. (2001). "Estimating variance of movements for time series data: issues arising for the Average Earnings Index". Unpublished report for the National Statistics Methodology Advisory Committee, ONS.