Journal of Official Statistics
Vol. 1, No. 3, 1985, pp. 315–322

Inverse Autocorrelations and Moving-Average Time Series Modelling

O. B. Oyetunji[1]

Abstract: Estimation of moving-average time series models has been difficult because it requires nonlinear optimization. Inverse autocorrelations, introduced by Cleveland (1972), offers a solution to this problem. We propose that moving-average modelling should use the techniques of the subset modelling method in which all possible subsets are evaluated. To illustrate this method we have used two sets of data, a simulated time series and Box and Jenkins (1970) Series A.

Key words: Autocorrelations; inverse autocorrelations; moving average model; subset model; BIC.

1. Introduction

Suppose we have a zero-mean stationary time series $\{X_t\}$. A moving-average (MA) model of order q, MA(q), is defined by

$$X_t = \varepsilon_t - \beta_1\varepsilon_{t-1} - \beta_2\varepsilon_{t-2} - \ldots - \beta_q\varepsilon_{t-q}\,, \qquad (1.1)$$

where $\{\varepsilon_t\}$ is a white-noise process with finite variance σ^2. The parameters are real constants such that their characteristic polynomial

$$B(z) = \sum_{j=0}^{q} \beta_j z^j, \quad \beta_0 = 1$$

has all its zeros outside the unit circle to ensure invertibility. Let $\gamma_k = \mathrm{E}(X_t\, X_{t+k})$ and $\rho_k = \gamma_k/\gamma_0$, $k=0,1,\ldots$ denote the autocovariance and autocorrelation functions of $\{X_t\}$, respectively. Then

$$\gamma_k = \begin{cases} \sigma^2(-\beta_k + \sum_{j=1}^{q-k} \beta_j\beta_{j+k}), \\ \qquad\qquad k = 0, \pm 1, \ldots, \pm q\,. \\ 0 \qquad\qquad |k| > q \end{cases} \qquad (1.2)$$

This means that the autocovariance (autocorrelation) function cuts off at lag $k = q$.

Many authors, Durbin (1959), Walker (1961), Hannan (1969), to mention a few, have proposed various efficient methods for estimating the parameters of a moving-average process. All such methods involve some form of nonlinear optimization. They are therefore difficult, cumbersome, and time-consuming to apply.

The moving-average process given by (1.1) is called a full-order model. We then define a subset MA process, like that of (1.1) where some of the β coefficients can be equal to zero, but β_q must be a nonzero number. We use the notation (ℓ)MA(q) to denote a subset MA process with ℓ nonzero coefficients and maximum lag q. Subset MA modelling has been difficult because there are many competing models. For a maximum specified lag q, there are 2^q possible models, and we have already pointed out the difficulty of fitting a single moving-average model.

[1] Department of Statistics, University of Ibadan, Ibadan, Nigeria.

However, Cleveland (1972) has introduced a new tool, inverse autocorrelations, which makes it possible to estimate MA process parameters by solving a set of prescribed simultaneous linear equations, as one does in autoregression. Subset modelling is another technique made possible by the emergence of fast electronic computers. McClave (1975) proposed a method that selects a subset autoregressive model without evaluating all the possible subsets from a maximum specified lag L. McClave's omission of viable and potentially viable subsets led Haggan and Oyetunji (1984) to propose a method of subset autoregression that evaluates all possible subsets. They showed that their algorithm is as fast as McClave's for $L \leqslant 15$, and sometimes faster, in identifying the best subset model. BIC, as described by Akaike (1977) and defined by formula 1.3, is the criterion used for identifying the best subset model. BIC for a q-variate model (a model with q nonzero coefficients) is defined as

$$\mathrm{BIC}(q) = (N-q)\ell n\tilde{\sigma}_q^2 - (N-q)\ell n\{1-(q/N)\} + q\ell n\{q^{-1}(\sigma_0^2-\tilde{\sigma}_q^2)+q\ell nN. \quad (1.3)$$

Here, N is the number of observations in the series, σ_0^2 is the data variance and $\tilde{\sigma}_q^2$ is the residual variance after fitting the q-variate model. BIC was defined specially for selection of subset models and Oyetunji (1979) has shown that among existing selection criteria, BIC's allround performance in selecting a subset surpasses all other criteria. For more details on order determination see Priestley (1981), Sections 5.4.5, 7.8, and 9.4.

It has been shown by Cleveland (1972) and Chatfield (1979) that the inverse autocorrelation function behaves for moving-average processes in exactly the same way as the autocorrelation function behaves for autoregressive processes. In this paper, we use the method of subset selection proposed by Haggan and Oyetunji (1984) to replace autocorrelations with inverse autocorrelations for the purpose of demonstrating that moving-average modelling can be as straightforward as autoregressive modelling.

In the search for the best subset moving-average model, using the method of evaluating all possible subsets from a specified maximum lag L, all full-order models up to lag L are evaluated. The selected subset model is as good as, if not better than the model we would have selected if we had considered full-order models only. Through combining modelling methods, we have arrived at a moving-average modelling technique that uses a form of subset selection in which all possible subsets are considered. Although we talk about subset selection, we also include full-order models. For example, one of our simulated series (used to describe the proposed method) is a full-order moving-average model, even though we identify it through subset selection.

We have already pointed out that Haggan and Oyetunji's method of evaluating all possible subsets is as fast as McClave's method for $L \leqslant 15$, although it explodes when $L > 15$ since there are 2^L possible subsets to evaluate. However, a specified maximum lag of 15 is large enough for most situations, and in particular, it is large enough to accommodate any seasonal variation in monthly data.

2. Inverse Autocorrelations

Suppose we have a stochastic process $\{X_t\}$ with spectral density function $f(\omega)$, autocovariance function $\gamma(k)$ and autocorrelation function $\rho(k)$, $k = 0,1,\ldots$. Then

$$\gamma(k) = \int_{-\pi}^{\pi} e^{i\omega k} f(\omega)\, d\omega,$$

and

$$\rho(k) = \gamma(k)/\gamma(0).$$

Let

$$fi(\omega) = 1/f(\omega). \tag{2.1}$$

Cleveland (1972) defined the inverse autocovariance function of $\{X_t\}$ as

$$\gamma i(k) = \int_{-\pi}^{\pi} e^{i\omega k} fi(\omega)\, d\omega, \tag{2.2}$$

and

$$\rho i(k) = \gamma i(k)/\gamma i(0), \tag{2.3}$$

is the inverse autocorrelation function.

Exploiting the duality between moving average and autoregressive processes, Cleveland shows that by using inverse autocorrelations, fitting a moving-average process is reduced to solving a set of prescribed simultaneous linear equations. However, estimating an inverse autocorrelation function is not as straightforward as estimating an autocorrelation function. We first have to estimate the spectral density function $f(\omega)$.

Although the idea of inverse autocorrelations is intuitively appealing, it did not gain immediate popularity. Chatfield (1979) suggested that a reason could be that Cleveland used a frequency-domain definition and that this form of expression made the idea of inverse autocorrelations difficult to grasp. Chatfield has, on the other hand, used a time-domain definition. However, recently there has been renewed interest in inverse autocorrelations. McClave (1978), Chatfield (1979), Hosking (1980), Bhansali (1980, 1983), and Battaglia (1983), to mention a few, have discussed estimation, asymptotic properties and use of inverse autocorrelations.

3. Estimation of Inverse Autocorrelations

Cleveland suggested two methods of estimating the inverse autocorrelation function; both methods stem from spectral density estimation techniques. The first method of estimating the spectral density function is to fit an autoregressive model using a high enough order to give a good fit. The problem with this method is that we have to impose a model on the series. The second method, which we have adopted in this paper, is to smooth the periodogram, $I(\omega)$, given by

$$I(\omega) = \frac{1}{2\pi N}\Big|\sum_{t=1}^{N}(X_t - \bar{X})\, e^{i\omega t}\Big|^2 = \frac{1}{2\pi}\Big[c(0) + 2\sum_{k=1}^{N-1} c(k)\cos\omega k\Big], \tag{3.1}$$

where $c(k)$ denotes the sample estimate of the autocovariance of lag k. Although the periodogram, $I(\omega)$, is asymptotically unbiased for the spectral density function, $f(\omega)$, its variance does not decrease as N increases. It is therefore necessary to smooth the periodogram, that is, apply some weighting function to $I(\omega)$. There are a number of weight functions, usually referred to as windows, that are commonly used. The weight function used here is the Daniell window and we estimate the inverse autocorrelations as follows. Let

$$\omega_j = \frac{2\pi j}{N},\ j = 0,1,\ldots,N-1\ .$$

Calculate

$$I(\omega_j) = \frac{1}{\pi}\Big[c(0) + 2\sum_{k=1}^{N-1} c(k)\cos\omega jk\Big],$$

and obtain an estimate of the spectral density by

$$\hat{f}(\omega_j) = \frac{1}{2m+1}\sum_{k=-m}^{m} I(\omega_{j+k}), \tag{3.2}$$

where m is a suitably chosen positive integer. We then estimate the inverse autocovariances by

$$ci(k) = \sum_{j=0}^{N-1} [e^{ik\omega j}/\hat{f}(\omega_j)] , \qquad (3.3)$$

and estimates of inverse autocorrelations are then obtained by

$$ri(k) = ci(k)/ci(0) . \qquad (3.4)$$

The main problem with this method is that there is no accurate way of choosing m. Hitherto, the choice of m has been purely a subjective process, most commonly done by plotting the smoothed periodogram for different values of m and choosing that m which gives the smoothest picture, without losing any characteristic feature of the spectrum. Bhausali (1980) has given a theoretical analysis of autoregressive and window estimates of inverse autocorrelation functions.

McClave (1978) has suggested another method of estimating the inverse autocorrelations in the frequency domain which first requires the fitting of a MA(q) process using one of the traditional methods. This method is therefore cumbersome and the advantage of computational savings is lost. Also, we need to impose an MA(q) model on the series before estimating its inverse autocorrelations.

Chatfield (1979) defined inverse autocorrelations in the time domain and gave a time-domain method of estimation. Although this method has some limitations, it has great appeal because of its ease. We discuss this method in Section 6.

4. Evaluating All Possible Subset Models

We have proposed a subset approach to moving-average modelling. In our approach, we use a method of evaluating all possible subset models that is given in Haggan and Oyetunji (1984), and use the BIC criterion for selecting the best model.

In Haggan and Oyetunji (1984), all possible subsets are evaluated. For each possible subset model, both the residual variance and the corresponding BIC are calculated. At each stage, only the model with the smallest BIC is noted together with the lags in that model. By the end of the iteration, the model with minimum BIC has been identified together with the lags in that model. Finally, coefficients of the lags in the model with minimum BIC are estimated by solving the corresponding Yule-Walker equations. Because more than one model may have similar values of BIC around the minimum, the method described above was modified so that three models with the smallest values of BIC are identified and estimated. For real data, the final model is selected by applying diagnostic checking (Box and Jenkins (1970)). Another possible diagnostic check that may be applied is to compare the parametric spectrum of the fitted models with the nonparametric spectrum of the raw data.

In this paper, we replace $c(k)$ by $ci(k)$ in the usual Yule-Walker equations to obtain the MA analogue, which we write in matrix form as

$$Ci_q = -\Gamma i_q \hat{\beta} , \qquad (4.1)$$

where

$$Ci_q = \begin{bmatrix} ci(1) \\ ci(2) \\ \vdots \\ \vdots \\ ci(q) \end{bmatrix} ,$$

$$\Gamma i_q = \begin{bmatrix} ci(0) & ci(1).....ci(q-1) \\ ci(1) & ci(0).....ci(q-1) \\ \vdots & \\ \vdots & \\ ci(q-1) & .....ci(0) \end{bmatrix} ,$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \vdots \\ \hat{\beta}_q \end{bmatrix} .$$

The augmented matrix to which Haggan and Oyetunji's algorithm is applied is then given by

$$C = \left[\begin{array}{c|c} \Gamma i_q & Ci_q \\ \hline Ci_q{}^T & ci(0) \end{array}\right].$$

From (4.1), estimates of the β's in a selected subset model are obtained by solving

$$\hat{\beta}^* = -\Gamma i_q{}^{*-1}\, Ci_q{}^*$$

where $\hat{\beta}^*$, $\Gamma i_q{}^*$ and $Ci_q{}^*$ are respectively $\hat{\beta}$, Γi_q and Ci_q with some elements constrained to zero (see Haggan and Oyetunji (1984)). An estimate of the residual variance of the fitted model is then

$$\hat{\sigma}^2 = \sigma_0^2/(1+\sum_{j=1}^{q} \hat{\beta}_j^2)\,,$$

where σ_0^2 is the data variance and some of the corresponding $\hat{\beta}_j$'s are zero.

5. Simulated Data

To demonstrate the feasibility of the method described above, we simulated two models:

Model 1: $X_t = \varepsilon_t - .6\varepsilon_{t-1} - .79\varepsilon_{t-2} + .504\varepsilon_{t-3}$

Model 2: $X_t = \varepsilon_t - .5\varepsilon_{t-4}$

where $\{\varepsilon_t\}$ is a normal white-noise process with unit variance. It should be noted that even though Model 1 is a full-order process, we identify it through subset modelling. We also point out that Model 2 is exactly the same as Model 2 in McClave (1978).

For each model, we generated 20 independent series, each series containing $N = 200$ observations. We then applied the method summarized in Section 4 to each series. For each series, we noted whether the correct model was chosen as the best, second-best, or third-best model ("best" in terms of minimum BIC). The results are summarized below.

	Best	2nd best	3rd best	Total
Model 1	17	1	2	20
Model 2	19	0	1	20

For Model 2, the result in McClave (1978) shows that his method identified the correct model 14 times out of 20 when $N = 200$. When $N = 1000$, he identified the correct model 20 times out of 20. Here, we have not bothered to generate series with $N = 1000$ because it is highly unlikely one would come across such a long series in practice. McClave's method suffers from its inability to identify more than the best model. Our method is easily adapted to identify more than one model around the minimum BIC, with the final model selected through diagnostic checks. For simulated data, however, diagnostic checking is not necessary since we know the exact model.

6. Time Domain Approach

Chatfield (1979) gave the time domain definition of the inverse autocorrelation function as follows. Let the autocovariance generating function for a time series $\{X_t\}$ be defined by

$$\Gamma(z) = \sum_{k=-\infty}^{\infty} \gamma(k) z^k\,, \tag{6.1}$$

where $\{\gamma(k)\}$ is the autocovariance function. The inverse autocovariance generating function $\Gamma I(z)$ is then defined by

$$\Gamma(z)\,\Gamma I(z) = 1. \tag{6.2}$$

The coefficient of z^k in the expansion of $\Gamma I(z)$ is called the inverse autocovariance of lag k

and denoted by $\gamma i(k)$. The inverse autocorrelation of lag k is then defined by

$$\rho i(k) = \gamma i(k)/\gamma i(0), \qquad k=0,1,\ldots .$$

Chatfield then used the idea of an inverse process to describe inverse autocorrelations. If a time series $\{X_t\}$ satisfies an autoregressive process of order p, AR(p), given by

$$X_t=\alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon_t, \tag{6.3}$$

where $\{\varepsilon_t\}$ is a white-noise process with variance σ_ε^2, the corresponding inverse process is the MA(p) given by

$$X_t = \varepsilon_t - \alpha_1\varepsilon_{t-1} - \ldots - \alpha_p\varepsilon_{t-p}. \tag{6.4}$$

The inverse autocorrelation function of the time series $\{X_t\}$ which satisfies the AR(p) of (6.3) is then the autocorrelation function of the MA(p) given by (6.4). That is,

$$\rho i(k) = \begin{cases} (-\alpha_k + \sum_{j=1}^{p-k} \alpha_j \alpha_{j+k})/(1+\sum_{j=1}^{p} \alpha_j^2), \\ \qquad\qquad k=\pm 1,\ldots,\pm p, \\ 0 \qquad\qquad |k| > p \end{cases} \tag{6.5}$$

and

$$\gamma i(0) = (1+\sum_{j=1}^{p} \alpha_j^2) / \sigma_\varepsilon^2 . \tag{6.6}$$

Thus, Chatfield's time domain method of estimation entails fitting an AR(p) to the time series $\{X_t\}$ using AIC according to Akaike (1974). Let $\hat\alpha_1, \hat\alpha_2, \ldots, \hat\alpha_p, \hat\sigma_\varepsilon^2$ denote estimates of parameters and the residual variance of the fitted model. Estimates of inverse autocorrelations of $\{X_t\}$ are then obtained by substituting $\hat\alpha_1, \hat\alpha_2, \ldots, \hat\alpha_p, \hat\sigma_\varepsilon^2$ in (6.5) and (6.6).

We applied this method to the 197 observations of Series A (Chemical Process Concentration Readings) given in Box and Jenkins (1970), p. 525. Box and Jenkins fitted an MA(1) to the first differences, given by

$$Y_t = \varepsilon_t + .70\varepsilon_{t-1} ,$$

where $Y_t = X_t - X_{t-1}$. Also, McClave (1978), using inverse autocorrelations, fitted the subset moving-average model to the first differences, given by

$$Y_t = \varepsilon_t - .647\varepsilon_{t-1} - .167\varepsilon_{t-4}.$$

First, we fitted an AR to the first differences. The AR selected by AIC is the AR(6) given by

$$Y_t + .60306Y_{t-1} + .39070Y_{t-2} + .35546Y_{t-3}$$

$$+ .31388Y_{t-4} + .31217Y_{t-5} + .21272Y_{t-6}$$

$$= \varepsilon_t, \quad \hat\sigma_\varepsilon^2 = .09635.$$

We then substituted the estimated coefficients and $\hat\sigma_\varepsilon^2$ in (6.5) and (6.6) to obtain the sample inverse autocorrelations, up to lag 6, for Y_t.

These estimates of inverse autocorrelations were then used in the subset selection algorithm of Haggan and Oyetunji. The best model, in terms of minimum BIC, is the MA(1) given by

$$Y_t = \varepsilon_t - .665\varepsilon_{t-1}, \text{ BIC} = 474.17, \tag{6.7}$$

while the second-best model is

$$Y_t = \varepsilon_t - .646\varepsilon_{t-1} - .057\varepsilon_{t-4}, \text{ BIC} = 477.87. \tag{6.8}$$

Thus, the best model coincides with Box and Jenkins MA(1), while the second-best model coincides with McClave's subset MA model. However, it should be pointed out that in arriving at his subset model, McClave did not

evaluate full-order models. The final model is selected through diagnostic checking. For example, the spectrum of (6.7) and the spectrum of (6.8) may be compared with the nonparametric spectrum of the raw data.

7. Summary and Conclusion

Our results from simulated data show conclusively that once the inverse autocorrelations have been estimated, the method of Haggan and Oyetunji (1984) can easily be adapted for subset moving-average selection. In our simulation of Model 2, we identified the correct model 19 times out 20 as the best model and once as the third-best model, whereas McClave's method identified it 14 times out of 20. Our method also has the advantage that it can be adapted to select a number of models having similar values of BIC around the minimum, with the final model being selected through diagnostic checks. In the search for the best subset model all the full-order models are also evaluated, so that if the correct model is a full-order model, it will be so identified through this subset selection method.

Bhansali (1983) has also pointed out that, compared to existing methods of fitting moving-average models, like those of Walker (1961), Hannan (1969) and Box and Jenkins (1970), the method of inverse autocorrelations has a number of advantages. It demands less computational effort and guarantees invertibility.

The inverse autocorrelation function is now receiving the prominence it deserves because it has made MA modelling easy. Not only this, it is proving to be more useful in model identification than the traditional partial autocorrelation function. Although both partial autocorrelations and inverse autocorrelations can be used to determine the order of AR to fit, Cleveland (1972) has shown that inverse autocorrelations will also indicate which lags have coefficients that are equal to zero. Cleveland pointed out that the inverse autocorrelations of the original 197 observations of Series A, Box and Jenkins (1970), suggested the model

$$X_t + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_7 X_{t-7} + \mu = \varepsilon_t.$$

After applying the subset autoregression method of Haggan and Oyetunji (1984) to Series A, the best model fitted was

$$X_t - .38X_{t-1} - .22X_{t-2} - .18X_{t-7} - 3.66 = \varepsilon_t \ ,$$

$$\hat{\sigma}^2 = .095 \ .$$

which agrees with Cleveland's suggestion. Box and Jenkins fitted

$$X_t - .92X_{t-1} = 1.45 + \varepsilon_t - .58\varepsilon_{t-1}, \ \hat{\sigma}^2 = .097.$$

In Section 6 we have shown that it is feasible to estimate inverse autocorrelations using Chatfield's time-domain method. The method is much easier and much faster than frequency-domain methods. However, when using Chatfield's method, the number of inverse autocorrelations that can be estimated is limited to the order of AR fitted. Although this time-domain estimation looks promising, there is still a lot of work to be done. As Chatfield himself pointed out, the properties of time-domain estimation still have to be investigated.

8. Note

In practice, energy data have been fitted by moving-average processes. For example, Brubacher and Tunnicliffe-Wilson (1976) have fitted a complex moving-average process to the differenced series of hourly total electricity demand in the province of Ontario, Canada. Application of the subset approach described in this paper would yield a simpler moving-average process.

Abraham and Ledolter (1984) have suggested that it may not be worthwhile estimating parameters in moving-average processes

using $ri(k)$ since maximum likelihood estimates are now readily available. Using simulated AR (2) data, they suggested that inverse autocorrelations may lead to underestimation of the order of an autoregressive process especially for small sample sizes. In this paper we have used neither partial autocorrelation nor inverse autocorrelation for order determination. We used AIC and BIC. However, using the inverse of Abraham and Ledolter's argument, one can suggest that partial autocorrelations may lead to overestimation of order. Maximum likelihood estimates are readily available but they are time consuming. We have shown that by using inverse autocorrelations, fast and efficient algorithms can be developed for fitting moving-average processes. Properties of such estimated parameters are already given in Bhansali (1980, 1983).

9. References

Abraham, B. and Ledolter, J. (1984): A Note on Inverse Autocorrelations. Biometrika, Vol. 71, pp. 609–614.

Akaike, H. (1974): A New Look at Statistical Model Identification. IEEE Trans, Aut. Contr. Vol. AC–19, No. 6, pp. 716–723.

Akaike, H. (1977): On Entropy Maximisation Principle. Proceedings from Symposium on Applications of Statistics, Ed. by P.K. Krishnaiah, North Holland, Amsterdam.

Battaglia, F. (1983): Inverse Autocovariances and Measure of Linear Determinism of a Stationary Process. Journal of Time Series Analysis, Vol. 4, No. 2, pp. 79–88.

Bhansali, R. J. (1980): Autoregressive and Window Estimates of the Inverse Correlation Function. Biometrika, Vol. 67, pp. 551–566.

Bhansali, R. J. (1983): Estimation of the Order of a Moving Average Model from Autoregression and Window Estimates of the Inverse Correlation Function. Journal of Time Series Analysis, Vol. 4, No. 3, pp. 137–162.

Brubacher, S. R. and Tunnicliffe–Wilson, G. (1976): Interpolating Time Series with Application to the Estimation of Holiday Effects on Electricity Demand. Applied Statistics, Vol. 25, pp. 107–116.

Box, G. E. P. and Jenkins, G. M. (1970): Time Series Analysis Forecasting and Control. Holden-Day, San Francisco.

Chatfield, C. (1979): Inverse Autocorrelations. Journal of the Royal Statistical Society, Series A, Vol. 142, pp. 363–377.

Cleveland, W. S. (1972): The Inverse Autocorrelations of a Time Series and Their Applications. Technometrics, Vol. 14, No. 2, pp. 277–297.

Durbin, J. (1959): Efficient Estimation of Parameters in Moving Average Models. Biometrika, Vol. 28, pp. 233–244.

Haggan, V. and Oyetunji, O. B. (1984): On the Selection of Subset Autoregressive Time Series Models. Journal of Time Series Analysis, Vol. 5, No. 2, pp. 103–113.

Hannan, E. J. (1969): The Estimation of Mixed Moving Average Autoregressive Systems. Biometrika, Vol. 56, pp. 579–593.

Hosking, J. R. M. (1980): The Asymptotic Distribution of the Sample Inverse Autocorrelations of an Autoregressive – Moving Average Process. Biometrika, Vol. 67, pp. 223–226.

McClave, J. (1975): Subset Autoregression. Technometrics, Vol. 17, No. 2, pp. 213–219.

McClave, J. (1978): Estimating the Order of Moving Average Models: the Max X^2 Method. Communications in Statistics, Vol. A 7(3), pp. 259–276.

Oyetunji, O. B. (1979): Ph. D. thesis, Department of Mathematics, University of Manchester, Institute of Science and Technology, U. K.

Priestley, M. B. (1981): Spectral Analysis and Time Series. Academic Press, London.

Walker, A. M. (1961): Large Sample Estimation of Parameters in Moving Average Models. Biometrika, Vol. 46, pp. 306–316.

Received Jan 1985
Revised April 1985

Journal of Official Statistics
Vol. 1, No. 3, 1985, pp. 323–329

Computing Methods for Variance Estimation in Complex Surveys

D. R. Bellhouse[1]

Abstract: A description is given of a computer program which calculates estimates of the variance – covariance matrix for estimates of means, totals and proportions at any stage of a multistage sampling design. The computer program uses tree traversal algorithms in which the sampling design structure is made equivalent to an unbalanced tree. Extensions to post-stratification and variance estimation for complex statistics are also discussed.

Key words: Multistage sampling; variance estimation; tree structures.

1. Introduction

Several computer programs have been developed to estimate standard errors of population estimates in sample surveys. Francis (1981) has given a summary of eleven of these programs. In some of the programs, for example CLUSTERS or SUPERCARP which are both described in Francis (1981), estimated standard errors or variances may be obtained for some specific sampling designs. In other programs, for example HES VAR X-TAB, described in Francis (1981), or subprograms in OSIRIS IV, described in Vinter (1980), the estimated variances for complex surveys are obtained by balanced repeated replication techniques. Thus, a survey researcher, when designing a survey in conjunction with these programs, is faced with one of two choices: choose a design which fits into one of the programs to obtain exact variance estimates, or choose a more general design and obtain approximate variance estimates. The computational technique described in this paper is a generalization of the researcher's first choice. It provides a method to compute exact variance estimates for general complex sampling designs based on the associated finite population sampling theory.

The computer program which implements these computational techniques is currently under development. To use this program, it is necessary only to provide the following information: the name of the sampling design and estimation procedure to be used at each stage, the size variable if a probability proportional to size sampling design has been used, the sample and population sizes, and the sample data in the appropriate order. The original method was described by Bellhouse (1980). A summary of this method is provided as well as extensions to post-stratification, to estimation of regression and other complex statistics, and to collapsed strata.

[1] Department of Statistical & Actuarial Sciences, University of Western Ontario, London, Ontario, Canada.

2. Variance-Covariance Estimation in General Multistage Designs

2.1. Sampling Theory Background

Consider a single-stage cluster sample of size n with sampled cluster totals $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ for two variables x and y. A linear estimator of the population total Y, of the variable y say, is $\hat{Y} = \sum_{i=1}^{n} w_i y_i$ where w_i, $i = 1, \ldots, n$, are weights either fixed in advance or determined from population and sampled auxiliary variables. The estimated covariance between $\hat{X}$ and $\hat{Y}$ may be described in general terms as $\text{cov}(\hat{X}, \hat{Y}) = g(\underset{\sim}{x}_s, \underset{\sim}{y}_s)$, a function of the sampled cluster totals, where $\underset{\sim}{x}_s = (x_1, \ldots, x_n)$, $\underset{\sim}{y}_s = (y_1, \ldots, y_n)$ and s denotes the sample. The estimated variance, $\text{var}(\hat{Y}) = g(\underset{\sim}{y}_s, \underset{\sim}{y}_s)$, is usually a quadratic form in $\underset{\sim}{y}_s$. Rao and Vijayan (1977) have obtained the necessary form of the nonnegative quadratic unbiased estimate of the variance, $\text{var}(\hat{Y})$. The covariance can be obtained by the standard technique of finding the variance of $\hat{D} = \hat{X} - \hat{Y}$.

Most of the standard unistage sampling estimators are linear in y. Their variances and covariances can be obtained from sampling texts such as Cochran (1977) or derived from the result of Rao and Vijayan (1977). The first step in the development of the computer program was to write FORTRAN subroutines which obtained estimates of means or totals and variance-covariance estimates for various unistage sampling designs and estimators. These subroutines include: simple random sampling using the sample mean or ratio estimator; sampling with probability proportional to size (pps) using the Horvitz-Thompson estimator with Sampford's (1967) design or the randomized pps systematic sampling design with joint inclusion probabilities given by Hidiroglou and Gray (1975); and cluster sampling using simple random sampling of clusters with either the unbiased estimator or the ratio estimator, or using probability proportional to the size of the cluster. The subroutine for unistage ratio estimation may be used for calculating the separate ratio estimator and its variance estimate. Theoretically, any pps sampling design could be used in this program in conjunction with the Horvitz-Thompson estimator. It is necessary only for the program user to write a subroutine which calculates the joint inclusion probabilities for the given sampling design.

Two-stage sampling variances and covariances may be obtained using the unistage subroutines. Raj (1966) and Rao (1976) have obtained general formulations of the variance of $\hat{Y}$ where the estimate $\hat{Y}$ is based on a two-stage sample. Bellhouse (1980) has given the associated covariances for each method. Both methods are of the form

$$\text{cov}(\hat{X}, \hat{Y}) = g(\hat{\underset{\sim}{x}}_s, \hat{\underset{\sim}{y}}_s) + \sum_{i=1}^{n} v_i \hat{c}_i , \tag{1}$$

based on estimates

$$\hat{X} = \sum_{i=1}^{n} w_i \hat{x}_i , \tag{2}$$

and

$$\hat{Y} = \sum_{i=1}^{n} w_i \hat{y}_i ,$$

where $g(\hat{\underset{\sim}{x}}_s, \hat{\underset{\sim}{y}}_s)$ is a copy of $g(\hat{\underset{\sim}{x}}_s, \hat{\underset{\sim}{y}}_s)$ with $\underset{\sim}{x}_s$ replaced by $\hat{\underset{\sim}{x}}_s$ and $\underset{\sim}{y}_s$ replaced by $\hat{\underset{\sim}{y}}_s$. The coefficients w_i and v_i, $i = 1, \ldots, n$, are known constants, and $\hat{c}_i$, is the estimated covariance between $\hat{x}_i$ and $\hat{y}_i$ within the sampled primary i, $i = 1, \ldots, n$. Stratified sampling is obtained upon setting $g(\hat{\underset{\sim}{x}}_s, \hat{\underset{\sim}{y}}_s) = 0$ in (1) and $n = N$ in the remaining term of (1) where N is the population size of primaries or the total number of strata.

This general formulation can be used recursively to obtain estimates and variance-covariance estimates for any multistage design. Consider three-stage sampling; the extension to four or more stages is straightforward. In this situation, a sample of primary units is obtained, then samples of secondary

units within each primary, and finally samples of tertiary units within each secondary. Begin at the final stage of sampling. Using the cluster sampling subroutines on the tertiary units, obtain estimates of the secondary totals or means and the associated variance-covariance estimates. Then go to the next stage up. Using formulae (1) and (2) with the estimates $\hat{\underset{\sim}{x}}_s$, $\hat{\underset{\sim}{y}}_s$ and $\hat{c}_i$ calculated from the previous stage, obtain estimates of the primary totals or means and the associated within primary variance-covariance estimates. Again, go to the next stage and repeat the same procedure. In this instance in formulae (1) and (2) $\hat{\underset{\sim}{x}}_s$ and $\hat{\underset{\sim}{y}}_s$ are the estimated primary totals and $\hat{c}_i$, $i = 1, \ldots, n$, are the estimated covariances within primaries.

One way to computerize this general estimation procedure is to impose a tree structure on the sampling design. In the traversal of the tree, all the appropriate calculations are made.

2.2. *Tree Structures and Multistage Designs*

The terminology used here for tree structures is that of Knuth (1968). For a k-stage sampling design a k-level tree is constructed; and for a k-stage sampling design with stratification a$(k+1)$-level tree is constructed. The tree will be unbalanced if there are unequal sample sizes at any stage of sampling other than the final. The nodes at the ith level in the tree contain the relevant information about the ith stage of sampling. The number of nodes at the ith level of the tree corresponds to the number of sampling units at the $(i-1)$th stage of sampling. Measurements on the sampled units at the final stage of sampling are stored in a separate data file appropriately ordered.

The method is illustrated by a simple example. Consider Data Set 2 given by Kaplan et al. (1979) to test the accuracy of the calculations performed by a number of sample survey package programs. The design used was stratified two-stage sampling with three strata. Within each stratum, a two-stage sample of three primaries was chosen with five units within each primary. The tree structure which corresponds to this sampling design is given in Fig. 1. Below the tree in Fig. 1 is the data file appropriately ordered. The vertical lines below the data values indicate the boundaries of the subsamples at the final stage of sampling. The tree in Fig. 1 is a balanced tree of three levels containing thirteen nodes labelled A, B_i $(i = 1, 2, 3)$ and C_{ij} $(i = 1, 2, 3;$ $j = 1, 2, 3)$. Node A is the root of the tree and nodes C_{ij} are terminal nodes.

The general tree construction algorithm used here has the following pattern. At any level in the tree, work from left to right. For each node in the current level, specify the number of nodes emanating from it to the next level. New storage locations for these lower level nodes and pointers to them are constructed. Then move to the next lower level. To construct the tree in Fig. 1, the number 3 is given to node A resulting in the creation of three storage locations for B_1, B_2, and B_3. Pointers to these storage locations are stored in A. The three branches from A to B_1, B_2, and B_3 correspond to the three strata in the design. The next step in the tree construction is to assign the number 3 to node B_1. Three new storage locations C_{11}, C_{12} and C_{13} are created and pointers to these locations are stored in B_1. The three branches in this subtree correspond to the three primary units chosen in stratum 1. Next, the number 3 is given to B_2 creating C_{21}, C_{22}, and C_{23} with the appropriate pointers in B_2. Finally, the number 3 is given to B_3 creating C_{31}, C_{32}, and C_{33}. At each step of the tree construction, additional information concerning the sampling design and desired estimator are given. At the root, node A, it is necessary to specify that the branches are strata. In each node $B_i(i = 1, 2, 3)$ the information given is that the design is simple random sampling of three primary units from a total of fifteen with the sample mean as the estimator. Finally, at

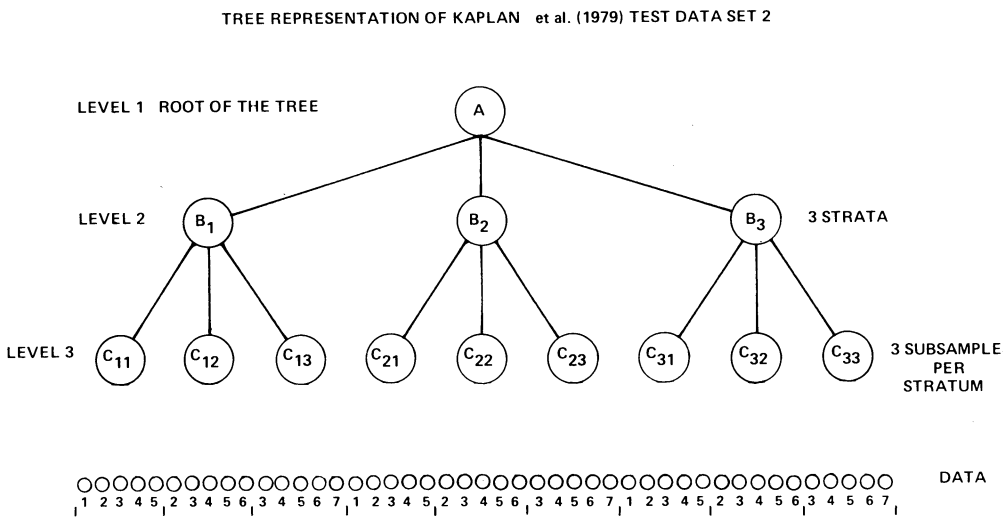

Fig. 1

the third level of the tree, the terminal nodes, C_{ij}, $i = 1, 2, 3$; $j = 1, 2, 3$ contain information about the "last stage of sampling," the subsampling of the primaries within the strata. In particular, the information given in each node is that the design is simple random sampling of five out of ten secondaries with the sample mean as the estimator. Finally, the data file that appears in Fig. 1 is in a specific order. The first five items belong to node C_{11}, i.e. they are the measurements on the five secondaries within the first primary in the first stratum. The next five items belong to node C_{12}, the following five to C_{13} and so on. Note that the data file is not part of the tree structure.

The type of traversal used in the program is endorder: the left-most subtree is traversed, then the second to left-most, and so on until the right-most subtree is traversed, and then the root of the subtree is visited. The algorithm used is a variation of one in Knuth (1968, pp. 317–319 and 560). At the time of construction of the tree, only the forward links exist. A pass could be made down the tree but not up. The backward links are created at the traversal stage. When a new node is reached in the traversal of the tree, a back pointer is given to its root. With endorder traversal, the nodes in Fig. 1 would be visited in the order $C_{11}C_{12}C_{13}B_1C_{21}C_{22}C_{23}B_2C_{31}C_{32}C_{33}B_3A$.

Assume that the population total is the item of interest. The calculations are performed as follows. Since the tree traversal is endorder, then the first node visited is C_{11}. The sample size of five indicates that five data points must be read from the data file, in this case the measurements 1, 2, 3, 4, 5. Using the information about the design and the estimate, a subroutine is called to calculate the estimate $\hat{Y} = 30$ and $v(\hat{Y}) = 25$. These values are stored in node B_1. The next node visited is C_{12} and the next five data items 2, 3, 4, 5, 6, are obtained from the file. The estimate $\hat{Y} = 40$ and $v(\hat{Y}) = 25$ are calculated from these data and stored in B_1. Then C_{13} is visited, data points 3, 4, 5, 6, 7 are read from the file and $\hat{Y} = 50$ and $v(\hat{Y}) = 25$ are stored in node B_1. Then node B_1 is visited. From the lower level of the tree or the lower stage in the design all the necessary data has been obtained. The estimate of the total for this primary is $\hat{Y} = 15(30 + 40 + 50)/3 = 600$ and the variance estimate

$v(\hat{Y}) = 15(15-3)[(30-40)^2 + (40-40)^2 + (50-40)^2]/[(3)\ (2)] + 15(25 + 25 + 25)/3 = 6375$. The values $\hat{Y} = 600$ and $v(\hat{Y}) = 6375$ are stored in node A. The next nodes visited are C_{21}, C_{22} and C_{23}, in that order, from which values $\hat{Y} = 30, 40$ and 50 respectively and $v(\hat{Y}) = 25, 25$ and 25 respectively are calculated and stored in B_2. Then B_2 is visited and the calculations $\hat{Y} = 600$ and $v(\hat{Y}) = 6375$ are made and stored in A. Then nodes C_{31}, C_{32} and C_{33}, in that order, are visited and $\hat{Y} = 30$, 40 and 50 respectively and $v(\hat{Y}) = 25$, 25 and 25 respectively are stored in node B_3. Then B_3 is visited and the calculation $\hat{Y} = 600$ and $v(\hat{Y}) = 6375$ are stored in A. Finally, node A is visited and $\hat{Y} = 600 + 600 + 600 = 1800$ and $v(\hat{Y}) = 6375 + 6375 + 6375 = 19125$ are calculated. The standard error of the estimate of the mean $(19125)^{1/2}/450 = .31$ agrees with the Kaplan et al. (1979) value.

The previous example is very simple and not indicative of typical survey data. Although the calculations for the preceeding example took only 3 seconds of CPU time on a PRIME 400 minicomputer, it remains to be seen whether the computing time would be excessive for larger and more complex surveys. Therefore the author obtained a larger data set which used a complex design. The data are from a survey of North American Indian children carried out in Canada during 1981–82. Six hundred responses with five variables each were analyzed. The Indians were divided into six strata by region of dwelling within Canada. Within a stratum, two, three or four enumeration areas (a Statistics Canada Census geographical area) were chosen by probability proportional to the size of the enumeration area. Sampford's (1967) design was assumed in the calculation of the joint inclusion probabilities. Within a chosen enumeration area, a number of families were chosen by simple random sampling and each child in a family was interviewed. The program produced both the estimates and estimated variance-covariance matrix. The calculations took 17 seconds of CPU time. The program also calculated and printed the estimates and variance-covariance estimates within each stratum so that interstratum comparisons could be made.

3. Post-Stratification

The method of calculating post-stratified variance estimates is based on the theory of Williams (1962). Suppose L post-strata are constructed. Let y denote the measurement on a sampling unit in the data file. Construct L new variables by setting $y_h = y$ if the sampling unit is in the hth post-stratum, 0 otherwise, $h = 1, \ldots, L$. Make one pass through the tree structure which defines the sampling design. During this pass, calculate an estimate of the population mean for each of the L data sets defined by the variables y_h, $h = 1, \ldots, L$. The resulting estimate $\hat{\bar{Y}}_h$ is the estimate of the mean in the post-stratum h, $h = 1, \ldots, L$. The post-stratified estimate is $\hat{\bar{Y}}_p = \sum_{h=1}^{L} W_h \hat{\bar{Y}}_h$, where W_h, $h = 1, \ldots, L$ are known stratum weights provided in advance. Now transform the original data points y by setting $x = y - \hat{\bar{Y}}_h$ if the sampling unit is the hth post-stratum, $h = 1, \ldots, L$. Then make a second pass through the tree structure. On this pass, calculate the estimated variance of $\hat{X}$, the estimated total based on the data x. The resulting estimate, $\text{var}(\hat{X})$ will be $\text{var}(\hat{\bar{Y}}_p)$, the post-stratified variance estimate of the estimated total $\hat{Y}_p = N\hat{\bar{Y}}_p$ for the data y, where N is the total population size. The estimated variance of $\hat{\bar{Y}}_p$, $\text{var}(\hat{\bar{Y}}_p) = \text{var}(\hat{Y}_p)/N^2$.

This method requires two passes through the data and the tree structure. However, only one set of operations by the program user is necessary: provide the stratum weights and the key words and numbers which describe the sampling design, the sample sizes, and other relevant information to perform the calculation.

4. Variance Estimates for the Simple Linear Regression Coefficient

The population regression coefficient may be expressed as

$$\hat{B} = \frac{\sum_{j=1}^{N} y_j(x_j - \bar{X})}{\sum_{j=1}^{N} (x_j - \bar{X})^2} ,$$

where N is the population size and the subscript j refers to an individual observation. This may be estimated by

$$\hat{B} = \frac{\sum_{j \epsilon s} w_j z_{1j}}{\sum_{j \epsilon s} w_j z_{2j}} , \qquad (3)$$

where s denotes the sample, unistage or multistage, w_j are the weights fixed in advance depending on the design, and where $z_{1j} = y_j(x_j - \hat{\bar{X}})$, $z_{2j} = (x_j - \hat{\bar{X}})^2$ and $\hat{\bar{X}} = \sum_{j \epsilon s} w_j x_j / \sum_{j \epsilon s} w_j$. Let the transformed variable

$$t_j = z_{1j} - \hat{B} z_{2j} . \qquad (4)$$

Three passes through the data and tree structure are necessary to calculate the estimated variance of $\hat{B}$. On the first pass, $\sum_{j \epsilon s} w_j x_j$ and $\sum_{j \epsilon s} w_j$, respectively, the estimated total for the x's and the estimated total for data which all have value 1, are calculated. After the second pass, $\hat{B}$ is calculated from (3). On this pass, both x and y are read from the data file and new variables z_1 and z_2 are derived. A one-pass algorithm could be derived to replace the first two passes. This would be analogous to the calculator and original formulae for sums of squares of deviations from a mean. As in this latter case, some numerical accuracy could be lost in the one-pass formulae. On the third pass through the tree structure and data file, calculate the estimated variance of $\hat{T}$, the estimated total based on the variable t from (4). Then $\text{var}(\hat{T})/(\sum_{j \epsilon s} w_j z_{2j})^2$ is the required variance estimate, where $\text{var}(\hat{T})$ is the variance estimator based on the derived variable t.

The program could also be adapted to compute variance estimates for other nonlinear statistics provided that the estimates can be expressed as functions of estimated totals. For example, both the separate and combined ratio estimators and their variance estimates can be obtained in one pass through the data and tree structure. For the combined ratio estimator, say $\hat{R}_c$, estimates $\hat{\bar{X}}$, $\hat{\bar{Y}}$, $\text{var}(\hat{\bar{X}})$, $\text{var}(\hat{\bar{Y}})$ and $\text{cov}(\hat{\bar{X}}, \hat{\bar{Y}})$ can be obtained in one pass. Then $\hat{R}_c = \hat{\bar{Y}}/\hat{\bar{X}}$ and $\text{var}(\hat{R}_c) = \{\text{var}(\hat{\bar{Y}}) - 2\hat{R}_c \text{cov}(\hat{\bar{X}}, \hat{\bar{Y}}) + \hat{R}_c^2 \text{var}(\hat{\bar{X}})\}/\bar{X}^2$. For the separate ratio estimator of the total, estimates of the subtotals and their estimated variances at the last stage of sampling are obtained by ratio estimation using the appropriate unistage subroutine. The tree traversal proceeds in the usual manner using these estimates of totals and variance estimates. Rao (1982) has given a review of variance estimation techniques for ratios, multiple regression and correlation coefficients based on the Taylor linearization method.

5. Estimation with One Unit per Stratum

When stratification has been carried out to the extent that there is only one unit per stratum, the method given in Cochran (1977) or the method of Hansen, Hurwitz and Madow (1953) utilizing an auxiliary variable may be used in the program to obtain variance estimates. Only one pass through the tree structure and data file is necessary. During this pass, only estimates of means or totals are obtained at each stage. The final node visited in the tree by endorder traversal is the root of the tree. When this node is reached, estimates

of the stratum totals will have been calculated and stored in this node. With the size of the groups and the auxiliary variable, if present, specified beforehand, the variance estimates are obtained using formulae (5A.56) or (5A.57) in Cochran (1977).

6. References

Bellhouse, D.R. (1980): Computation of Variance-Covariance Estimates for General Multistage Sampling Designs. COMPSTAT 1980: Proceedings in Computational Statistics, pp. 57–63, Physica-Verlag, Vienna.

Cochran, W.G. (1977): Sampling Techniques. 3rd Ed, Wiley, New York.

Francis, I. (1981): Statistical Software: A Comparative Review. North Holland, Amsterdam.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953): Sample Survey Methods and Theory. Vol 1, Wiley, New York.

Hidiroglou, M.A. and Gray, G.B. (1975): A Computer Algorithm for Joint Probabilities of Selection. Survey Methodology 1, pp. 99–108.

Kaplan, B., Francis, I. and Sedransk, J. (1979): Criteria for Comparing Programs for Computing Variances of Estimators from Complex Surveys: Proceedings of the 12th Interface Symposium, Waterloo, pp. 390–395.

Knuth, D.E. (1968): The Art of Computer Programming, Vol. 1. Addison-Wesley, Reading, Massachusetts.

Raj, D. (1966): Some Remarks on a Simple Procedure of Sampling without Replacement. Journal of the American Statistical Association, 61, pp. 391–396.

Rao, J.N.K. (1976): Unbiased Variance Estimation for Multistage Designs. Sankhya C, 37, pp. 133–139.

Rao, J.N.K. (1982): Some Aspects of Variance Estimation in Sample Surveys. Utilitas Mathematica, 21B, pp. 205–226.

Rao, J.N.K. and Vijayan, K. (1977): On Estimating the Variance in Sampling with Probability Proportional to Aggregate Size. Journal of the American Statistical Association, 72, pp. 579–584.

Sampford, M.R. (1967): On Sampling without Replacement with Unequal Probabilities of Selection. Biometrika 54, pp. 499–513.

Vinter, S. (1980): Survey Sampling Errors with OSIRIS IV. COMPSTAT 1980: Proceedings in Computational Statistics, pp. 72–80. Physica-Verlag, Vienna.

Williams, W.H. (1962): The Variance of an Estimator with Post-Stratified Weighting. Journal of the American Statistical Association, 57, pp. 522–627.

Received September 1984
Revised May 1985