Proceedings of the 1993 Winter Simulation Conference
G.W. Evans, M. Mollaghasemi, E.C. Russell, W.E. Biles (eds.)

INTERLACED VARIANCE ESTIMATORS

Demet Ceylan
Bruce W. Schmeiser

School of Industrial Engineering
Purdue University
West Lafayette, IN 47907-1287, U.S.A.

ABSTRACT

The sample variance, the usual estimator of the population variance, is biased when the data are autocorrelated. We investigate interlaced variance estimators, the average of k sample variances, each obtained using only every k^{th} observation; the sample variance is the special case of $k=1$. We analyze performance as a function of k for independent data, AR(1) processes, and MA(q) processes.

Interlacing reduces bias, in some cases to zero. Variance and mean squared error (mse) are asymptotically not a function of k. Because it is computationally simpler, the sample variance ($k=1$) is a reasonable choice for most applications.

1. INTRODUCTION

We consider estimating the population variance from a stationary time series. The usual estimator of the population variance, σ^2, is the sample variance,

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2 ,$$

where $X_1, X_2, \ldots, X_n$ are the n observations and $\bar{X}$ is the sample mean.

Ramberg, Sanchez, Sanchez and Hollick (1991), propose an alternative to the sample variance. They use every k^{th} observation to reduce the effect of autocorrelation. (A similar idea is used in the context of quantile estimation by Heidelberger and Lewis (1983) to obtain less correlated or ideally uncorrelated series.) This estimator, which we refer to as the interlaced variance estimator, is

$$\bar{S}^2 = \frac{1}{k}\sum_{j=1}^{k} S_j^2 ,$$

where $S_j^2 = \sum_{i=0}^{m-1}(X_{j+ik} - \bar{X}_j)^2 / (m-1)$ and $\bar{X}_j = \sum_{i=0}^{m-1} X_{j+ik} / m$ for $j=1,2,\ldots,k$.

In the following two sections we analyze the bias and variance, respectively, of the interlaced variance estimator, as a function of k. Section 4 is a summary.

2. BIAS OF THE INTERLACED VARIANCE ESTIMATOR

Bias, a measure of the accuracy of an estimator, is the difference between the expected value of the estimator and the population value. In this section we state the expected value of the estimator for identically independently distributed, (iid) data, MA(q), and AR(1) processes as a function of k. Interlacing does indeed reduce bias.

The expected value of the interlaced estimator is

$$\mathrm{E}\left(\bar{S}^2\right) = \mathrm{E}\left(S_j^2\right) = \sigma^2\left(1 - \frac{2}{m-1}\sum_{h=1}^{m-1}\left(1-\frac{h}{m}\right)\rho_{kh}\right),$$

or equivalently (Schmeiser 1990, p.314),

$$= \frac{m}{m-1}\left(\sigma^2 - \mathrm{V}\left(\bar{X}_j\right)\right)$$

where $\rho_h = \mathrm{Cov}\left(X_1, X_{1+h}\right) / \sigma^2$, $h=1,2,\ldots,n-1$, is the h-lag autocorrelation. The estimator is unbiased for q-dependent data if $k > q$, and for iid data for any positive integer k. That is, for any MA(q) model, $\mathrm{E}\left(\bar{S}^2\right) = \sigma^2$ if $k > q$.

Unlike MA(q) data, autoregressive processes do not have cut-off points for covariance. As expected, for AR(1) models increasing k always decreases the bias. We numerically showed that when the lag-1 correlation is 0.3, a skip of $k=10$ reduces bias 99 percent, essentially to zero. For lag-1 correlation of 0.9, a skip of $k=10$ reduces bias by roughly 50%.

In the next section we see that variance decreases inversely with sample size, as bias does in this section. Therefore, because mse is bias squared plus variance, bias is an asymptotically negligible component of mse.

3. VARIANCE OF THE INTERLACED VARIANCE ESTIMATOR

We analyze the variance of $\bar{S}^2$ for independent data and MA(q) and AR(1) processes.

For independent data (e.g., Wilks 1962)

$$V\left(\bar{S}^2\right) = \frac{\sigma^4}{n}\left(\alpha_4 - \frac{n-3k}{n-k}\right),$$

which is minimized when $k=1$. Hence the classical sample variance has minimal mse among interlaced variance estimators for iid data., but asymptotically the variance is not a function of k.

For MA(q) models with $k > q$

$$nV\left(\bar{S}^2\right) = \sigma^4\left(\alpha_4 - 1\right)$$

$$+ 2\sum_{h=1}^{q}\left(\mathrm{Cov}\left((X_1-\mu)^2,(X_{1+h}-\mu)^2\right)\right) + O(\frac{1}{m}).$$

Therefore, variance asymptotically does not depend on k. For finite sample sizes and MA(1) data an exact calculation shows that $k=1$ produces an estimator with variance that is negligibly less than for $k=2$.

For MA(q) normal processes, the correlation of squared observations is the correlation squared, and therefore for any fixed value of k

$$\lim_{n\to\infty} nV\left(\bar{S}^2\right) = 2\sigma^4\left(1+2\sum_{h=1}^{q}\rho_i^2\right).$$

This is also the asymptotic variance of the sample variance, S^2, for these processes. (Ceylan 1993). Therefore, as in the iid case, the asymptotic mse is not a function of k.

Our numerical analysis for AR(1) processes suggests that $k=1$ yields the estimator with the smallest variance for finite sample sizes. The choice of k has less effect on the variance of the estimator as the sample size increases. We proved that choice of k has no effect on the variance of the estimator in the limit, as $n \to \infty$.

4. CONCLUSION

We have analyzed the interlacing method for sample variances on several time-series models with various covariance structures. In each case, the asymptotic mse is not a function of k, the number of groups.

1) When the data are iid the interlaced variance estimator is unbiased for all k and the minimal variance is achieved with the usual estimator, $k=1$.

2) For MA(q) models, an unbiased estimator is obtained using $k > q$.

3) For AR(1) models, k can be increased to decrease the bias with a corresponding increase in variance. For the AR(1) examples that we studied, increasing k did not improve mse.

Therefore, interlacing is worthwhile only if bias is a primary concern. Those interested in a good (in the mse sense) estimator. can choose any value of k.

Computational complexity is a reason to choose $k=1$. The computational complexity and the memory requirement of interlacing increase linearly with k. In addition, coding is somewhat more complicated since $2k$ accumulators (for sums and sums of squares) must be maintained for the k sample variances, and these k sample variance must then be averaged.

REFERENCES

Ceylan, D. 1993. Variances in Dynamic-System Performance: Point Estimation and Standard Errors Ph.D. Preliminary Proposal, School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

Heidelberger, P. and Lewis, P.A.W. 1984. Quantile Estimation in Dependent Sequences. *Operations Research* 32:185-209.

Ramberg, J. S., Sanchez, S. M., Sanchez, P. J. and Hollick, L. J. 1991. Designing Simulation Experiments: Taguchi Methods and Response Surface Metamodels. In *Proceedings of the 1991 Winter Simulation Conference, ed.* B. L. Nelson, W. D. Kelton, G. M. Clark, 167-176.

Schmeiser, B. 1990. Simulation Experiments. In *Handbooks in OR & MS,* Vol 2 (D.P. Heyman, M. J. Sobel, eds). North-Holland: Elsevier Science Publishers B.V.

Wilks, S.S. 1962) *Mathematical Statistics*, New York: John Wiley

AUTHOR BIOGRAPHIES

DEMET CEYLAN is a Ph.D. student in the School of Industrial Engineering at Purdue University. She received B.S. and M.S. degrees in Industrial Engineering from Middle East Technical University in Turkey. Her research interests include simulation output analysis.

BRUCE SCHMEISER is a Professor in the School of Industrial Engineering at Purdue University. He is the Simulation Area Editor of *Operations Research* and an active participant in the Winter Simulation Conference, including being Program Chairman in 1983 and Chairman of the Board of Directors during 1988-1990.