Proceedings of the 1983
Winter Simulation Conference
S. Roberts, J. Banks, B. Schmeiser (eds.)

CONTROL VARIATES IN NONLINEAR REGRESSION

James J. Swain
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georga 30332

Control variates can be applied to Monte Carlo sampling experiments to improve the precision of the results. This method is especially useful in statistical problems were low order approximators of a particular variate of interest are available and possibly several statistical properties of the variate are to be investigated. In this paper a control variate scheme based on the linear approximator δ of the nonlinear parameter estimator $\hat{\theta}$ is used to improve the precision of the first four moments of $\hat{\theta}$ and the covariance matrix of the paramter estimates. The control variate method is shown to improve the effectiveness of the Monte Carlo results without substantially increasing the estimation effort, and it is effective over a wide range of nonlinearities. An approximate expression for the effectiveness of the control variate method based on the Beale measure of nonlinearity N_θ is given.

1. INTRODUCTION

In the nonlinear parameter estimation or regression problem, the p-vector of parameters $\underline{\theta}$ are estimated from the n responses y_i (i=1, ..., n) which are assumed to consist of the true response $\eta(\underline{x}_i ; \underline{\theta}_0)$ plus an additive error ε_i,

$$y_i = \eta(\underline{x}_i; \underline{\theta}_0) + \varepsilon_i. \qquad (1)$$

The ε_i are assumed to be independently and normally distributed random errors with variance σ^2, and $\eta(\underline{x}; \underline{\theta})$ is a nonlinear function of the parameters $\underline{\theta}$. The unknown parameters $\underline{\theta}_0$ are most often estimated by the method of least squares. That is, the estimator $\hat{\underline{\theta}}$ is chosen to minimize the sum of the squared residuals

$$S(\underline{\theta}) = \sum_{i=1}^{n} (y_i - \eta(\underline{x}_i; \underline{\theta}))^2 = \sum_{i=1}^{n} e_i^2(\underline{\theta}) \qquad (2)$$

so that $s(\hat{\underline{\theta}}) = \min S(\underline{\theta})$. Of course, other estimators for $\underline{\theta}$ can be used, but the least squares method is the most common and will be treated here.

The nonlinear regression problem is similar to that of linear regression, save that the nonlinearity of the response function complicates the numerical problem of obtaining the estimates and of resolving the statistical properties of the estimators obtained. The solution to the least squares problem is generally not a closed form function of the observed responses y_i so that iterative procedures are needed to obtain a solution (Bard, 1974; Gallant, 1975). The lack of a closed form solution makes the distribution of the estimator $\hat{\underline{\theta}}$ intractible, and the sampling distribution of the estimator $\hat{\underline{\theta}}$ is known exactly only asymptotically. The asymptotic approximation of the sampling distribution can be very misleading in problems with a small sample size. In particular, the finite sample estimator is almost always biased and the shape of the confidence region can differ markedly from the elliptical contours of the asymptotic sample distribution.

CH1953-9/83/0000-0623 $01.00 © 1983 IEEE

The statistical properties of the estimator $\hat{\underline{\theta}}$ can be approached using the asymptotic distribution, through a series approximation solution, or via Monte Carlo sampling. Under certain circumstances (see Gallant, 1975, for instance) the distribution of $\hat{\underline{\theta}}$ is asymptotically normal with mean $\underline{\theta}_0$ and variance matrix $(F^T(\underline{\theta}_0)F(\underline{\theta}_0))^{-1}\sigma^2$, where $F(\underline{\theta}_0)$ is the n by p. Jacobian matrix of first derivatives of $\eta(\underline{x};\underline{\theta})$ with respect to the parameters $\underline{\theta}$ and evaluated at $\underline{\theta}_0$. Note that the variance matrix is the limiting matrix as the sample size n increases indefinitely. To use the asymptotic approximation in practice one uses the finite sample and the two formulas given above. This solution is essentially the first order approximation to be described.

Series approximations can be obtained by approximating the sum of squares function $S(\underline{\theta})$ or its derivative $\partial S(\underline{\theta})/\partial\underline{\theta}$ (which vanishes at $\hat{\underline{\theta}}$) at some point such as $\underline{\theta}_0$. The approximator results from minimizing the approximate sum of squares function or solving for the point which solves the equation $\partial S(\underline{\theta})/\partial\underline{\theta} = 0$. For instance, the first order approximator $\underline{\delta}$ for $\hat{\underline{\theta}}$ is given by

$$\underline{\delta} = \underline{\theta}_0 + (F^T(\underline{\theta}_0)\, F(\underline{\theta}_0))^{-1}F^T(\underline{\theta}_0)\underline{\varepsilon}. \qquad (3)$$

This is similar to the asymptotic solution, in that $\underline{\delta}$ is normally distributed with mean $\underline{\theta}_0$ and variance matrix $(F^T(\underline{\theta}_0)\, F(\underline{\theta}_0))^{-1}\sigma^2$, although in this case the number of rows in $F(\underline{\theta})$ will be finite. The approximator $\underline{\delta}$ corresponds to the Gauss approximation method for finding a solution to the nonlinear least squares problem. Higher order approximations can be derived (Box, 1971; Clarke, 1981), but simplifications are introduced in each to make the solution tractable in the multiple parameter case. Note that measures of nonlinearity (Beale, 1960; Bates and Watts, 1980) such as Beale's N_θ have been developed to assess the extent to which $\underline{\delta}$ can be used to approximate the statistical properties of $\hat{\underline{\theta}}$. The measures can readily be adopted to any parameterization of the model, e.g., $\underline{\psi} = \underline{\psi}(\underline{\theta})$, and Beale denotes the least possible nonlinearity under any parameterization as the intrinsic nonlinearity, N_ϕ.

Monte Carlo methods can in principle be used to overcome the limitations of aproximation methods. A direct Monte Carlo algorithm consists of N repeated independent samples of the error vector $\underline{\varepsilon}(\underline{\varepsilon}_v,\ v=1,1,\ldots,N)$, from which (using (1)) the observed vectors $\underline{y}_v$ are obtained and used to obtain estimates $\hat{\underline{\theta}}_v$ by repeated solution of equation (2) for each $\underline{y}_v$. The N random sample points $\hat{\underline{\theta}}_v$ can be combined to obtain the sample statistics to estimate the properties of $\hat{\underline{\theta}}$. For instance, the kth marginal moments of $\hat{\underline{\theta}}$, $\underline{\mu}_k = E(\hat{\underline{\theta}})^k$ (where the exponentiation is component by component) can be estimated using the statistics $\hat{\underline{\mu}}_k$,

$$\hat{\underline{\mu}}_k = N^{-1}\sum_{v=1}^{N}(\hat{\underline{\theta}}_v)^k \qquad (4)$$

and the variance matrix of the parameter estimators $\hat{\underline{\theta}}$, $\mathrm{Var}(\hat{\underline{\theta}}) = \Sigma_\theta$ can be estimated using the sample statistic $\hat{\Sigma}_\theta$,

$$\hat{\Sigma}_\theta = (N-1)^{-1}\sum_{v=1}^{N}(\hat{\underline{\theta}}_v - \hat{\underline{\mu}}_1)(\hat{\underline{\theta}}_v - \hat{\underline{\mu}}_1)^T. \qquad (5)$$

The Monte Carlo method is conceptually simple and can be easily extended to nonlinear least estimators other than nonlinear least squares or to errors other than normal, but suffers from the drawback that N nonlinear estimation problems must be solved and that the standard errors of statistics such as $\hat{\underline{\mu}}_1$ decrease only as $N^{-1/2}$. Therefore, although arbitrary accuracy can be obtained via Monte Carlo sampling, arbitrarily many estimation problems may have to be solved to obtain that accuracy.

2. IMPROVED MONTE CARLO USING CONTROL VARIATES

The efficiency of Monte Carlo sampling can be improved by variance reduction techniques (Hammersley and Handscomb, 1964; Law and Kelton, 1982). These methods can take several forms, among which the best known are: antithetic variates, stratification, conditional expectations, and control variates. Control variates, using the linear approximator $\underline{\delta}$ (equation (3)) as the control variate is a very natural approach. This method is very simple to

implement, and it leaves the sampling method unchanged so that the sample estimates $\hat{\underline{\theta}}_v$ are all independent. The distribution of $\underline{\delta}$ is known and $\underline{\delta}$ is easier to compute than $\hat{\underline{\theta}}$ so that very little additional work is required to implement the method. In particular, the control $\underline{\delta}$ can be used as the basis for controls for all the marginal moments and covariances, as well as the quantiles. It is more efficient than direct Monte Carlo at all the levels of nonlinearity commonly encountered in practice.

The details of the control variate method are given in Swain (1982) and Swain and Schmesier (1983). The control estimators for the marginal moments $\underline{\mu}_k$, are given by

$$\hat{\underline{\mu}}_k(B_k) = \hat{\underline{\mu}}_k - B_k(\hat{\underline{\delta}}_k - E\hat{\underline{\delta}}_k) \qquad (6)$$

where

$$\hat{\underline{\delta}}_k = N^{-1} \sum_{v=1}^{N} (\underline{\delta}_v)^k$$

is the sample kth moment for $\underline{\delta}$ and the p by p matrix B_k is the control weight matrix. The method depends upon the covariation between $\hat{\underline{\theta}}_v$ and δ_v (they both depend upon the same vector of errors, $\underline{\varepsilon}_v$), so that when $\hat{\underline{\delta}}_k$ is greater than its expectation, it can be assumed that $\hat{\underline{\mu}}_k$ exceeds its expectation as well, and a correction proportional to $-(\hat{\underline{\delta}}_k - E\hat{\underline{\delta}}_k)$ should be made to $\hat{\underline{\mu}}_k$. It can be shown that the variance of $\hat{\underline{\mu}}_k(B_k)$ is minimized for the choice

$$B_k^* = \mathrm{Cov}((\hat{\underline{\theta}})^k , (\underline{\delta})^k) \, \mathrm{Var}^{-1}((\underline{\delta})^k),$$

and that the decrease in the control variate variance is largest when the correlation between $\underline{\delta}$ and $\hat{\underline{\theta}}$ is greatest. The covariance term in equation (7) for B_k^* is generally unknown, so B_k^* must be specified in some other way. Experience with the method has shown that the choice B_k = I (I is the p by p identity matrix) is nearly optimal and leads to a simple, unbiased estimator for $\underline{\mu}_k$. This is also consistent with the asymptotic case, since B_k^*(Eqn(7)) tends to the identity matrix as the sample size increases indefinitely.

A similar control variate scheme can also be given for Σ_θ. Let $\underline{S}_\theta$ be a vector of length m=p(p+1)/2 containing the lower diagonal elements of $\hat{\Sigma}_\theta$ stored by row (e.g., $\hat{\sigma}_{11}$, $\hat{\sigma}_{12}$, ... , $\hat{\sigma}_{pp}$). Let $\underline{S}_\delta$ be a similar m-vector for Σ_δ. Then

$$\underline{S}_\theta(C) = \underline{S}_\theta - C(\underline{S}_\delta - E(\underline{S}_\delta)) \qquad (8)$$

is a control variate estimator for Σ_θ and again by experience the choice of the m by m identity matrix for C is nearly optimal.

3. RESULTS

The efficiency of two Monte Carlo procedures can be computed as the ratio of their precisions, with precision measured as the inverse of the variance. For the multiple parameter case the scalar measure adopted is the determinant of the variance matrices, which is also known as the generalized variance. Then the efficiency of a control variate estimator for the kth moment compared to the crude estimator of the kth moment is given by

$$E_k = (1/ | \mathrm{Var}\, \hat{\underline{\mu}}_k(I)|)/(1/ | \mathrm{Var}\, \hat{\underline{\mu}}_k(0)|)$$
$$= | \mathrm{Var}\, \hat{\underline{\mu}}_k(0)|/ | \mathrm{Var}\, \hat{\underline{\mu}}_k(I)|$$

since direct Monte Carlo corresponds to the use of a control variates with a 0 weighting matrix. Note that efficiencies in excess of 1 indicate that the control variance is less than that of the direct Monte Carlo estimator.

The sample efficiencies (k=1) are plotted in figure 1 versus the nonlinearity measure N_θ, showing that the control variate estimator is more efficient than direct Monte Carlo for a wide range of nonlinearities, and the efficiency becomes infinite as the nonlinear estimator approaches the behavior of a linear estimator (i.e., asymptotically). The range of N_θ depicted is typical of the values encountered in practice.

The apparent relation between E_1 and N_θ is not surprising, since the Beale measure N_θ explicitly measures the appropriateness of $\underline{\delta}$ as

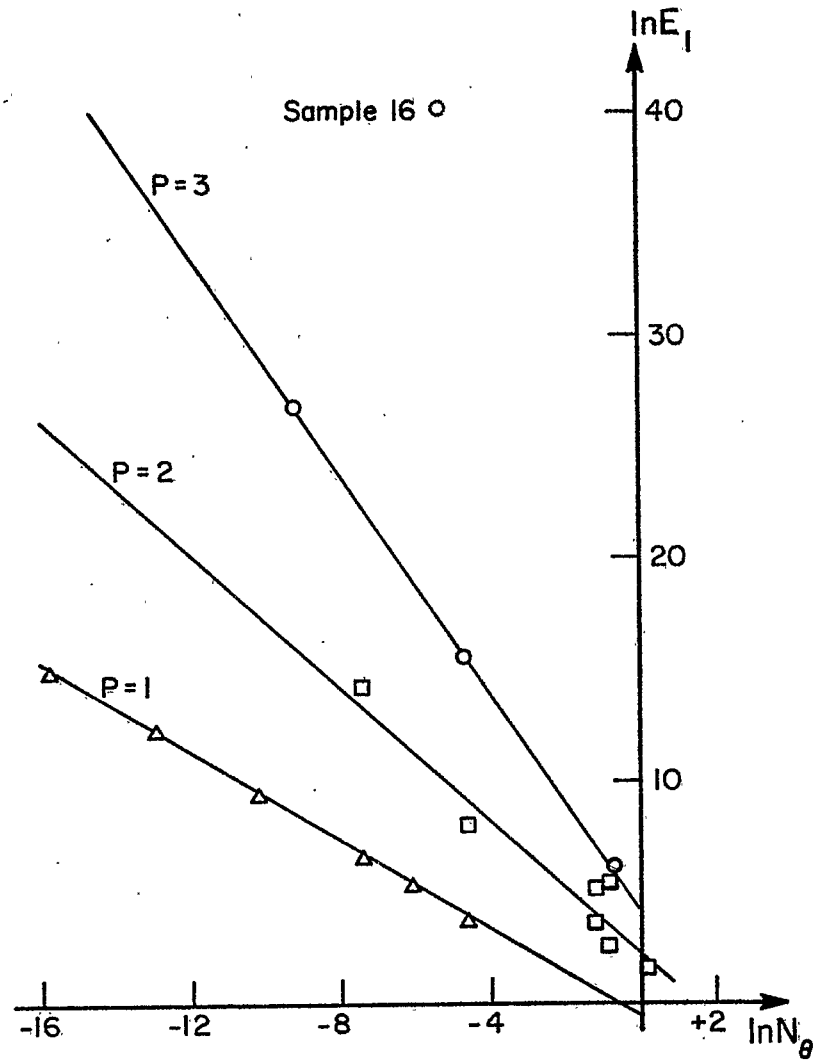

Figure 1. Summary of sampling efficiencies for the control estimator $\hat{\mu}_1(I)$ for 25 samples taken from Swain (1982). Some points are multiple samples. Sample 16 has one parameter which behaves like a linear estimator and therefore has higher than expected efficiency.

an approximator for $\hat{\theta}$ and δ is the control variate used here. Moreover, an approximate relation between the two (for k=1) is given (Swain, 1982) by

$$E_1 = (2N_\theta)^{-p} \qquad (9)$$

and the degree of this fit is given in figure 2.

The relation between efficiency and N_θ, equation (9), allows a prediction in advance of sampling of how efficient the control variate scheme is likely to be. However, since the control variable strategy is efficient for all but the most extreme nonlinearities (N_θ in excess of $1/F_{p,n-p;\alpha}$ can be considered extreme, according to Beale) and because the method requires only relatively simple linear computations, it will always be advantageous to use the method. In addition, equation (9) suggests that the efficiency of the method can be further improved by a suitable choice of a parameter transformation. An upper bound on the possible efficiency using control variables can be given by substituting the intrinsic

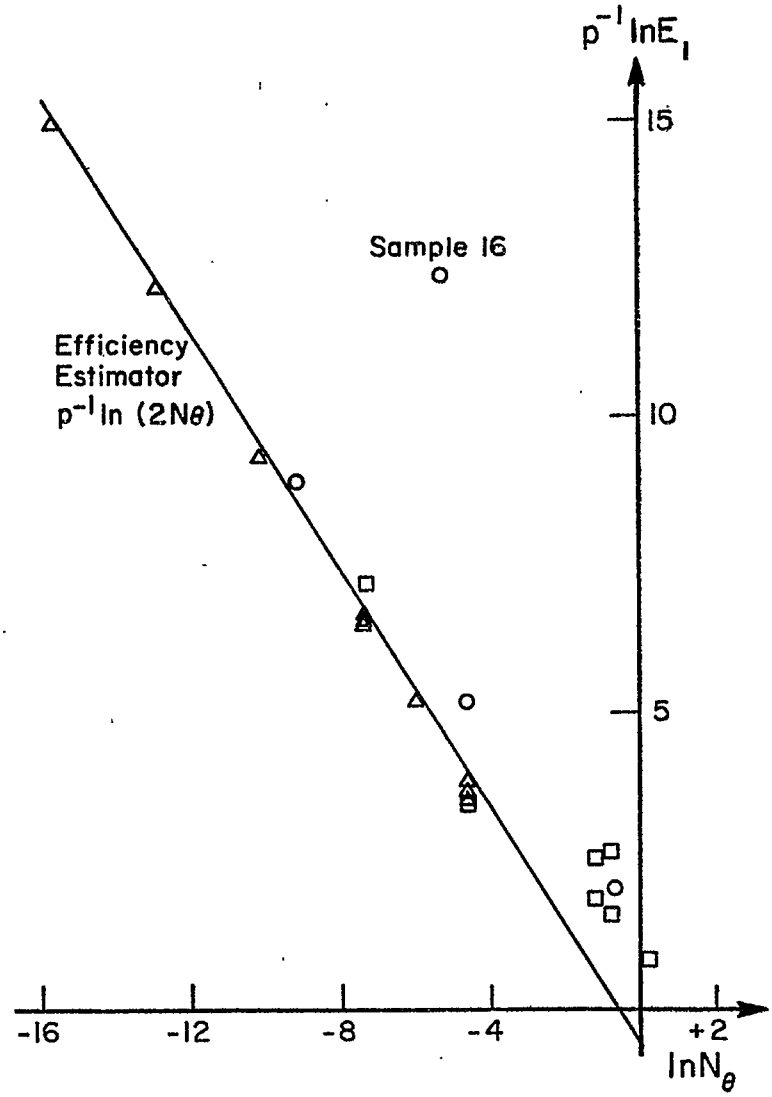

Figure 2. Approximate efficiency formula for the control estimator $\hat{\underline{\mu}}_1(I)$, $p^{-1}\ln E_1 = -\ln 2N_\theta$. The 25 samples are taken from Swain (1982).

nonlinearity N_ϕ in place of N_θ in equation (9). For instance, in certain cases a linear approximation based upon the transformed parameters $\underline{\psi} = (\underline{\theta})^k$ leads to additional efficiency in the control variate estimator for the marginal moments (Swain, 1982).

REFERENCES

Bard, Y. (1974), Nonlinear Parameter Estimation, Academic, New York.

Bates, D.M., Watts, D.G. (1980), Relative Curvature Measures of Nonlinearity, J. Royal Statistiacal Society, Series B, Vol. 42, No. 1, pp. 1-25.

Beale, E.M.L. (1960), Confidence Regions in Non-linear Estimation, J. Royal Statistical Society, Series B, Vol 22, No. 1, pp. 41-88.

Box, M.J. (1971), Bias in Nonlinear Estimation, J. Royal Statistical Society, Series B, Vol. 33, No. 1, pp. 171-201.

Clarke, G.P.Y. (1980), Moments of the Least Squares Estimators in a Non-linear Model, J. Royal Statistical Society, Series B, Vol. 42, No. 2, pp. 227-237.

Gallant, A.R. (1975), Nonlinear Regression, The American Statistician, Vol. 29, No. 2, pp. 73-81.

Hammersley, J.M., Handscomb, D.C. (1964), Monte Carlo Methods, Chapman and Hall, London.

Law, A.M., Keltyon, W.D. (1982), Simulation Modeling and Analysis, McGraw-Hill, New York.

Swain, J.J. (1982), Monte Carlo Estimation of the Sampling Distribution of Nonlinear Parameter Estimators, Unpublished Ph.D. Dissertation, Purdue University, West Lafayette, Indiana.

Swain, J.J., Schmeiser, B.W. (1983), Monte Carlo Estimation of the Sampling Distribution of Nonlinear Model Parameter Estimators, Technical Report C-83-1, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 17pp. (Also Purdue University Industrial Engineering Technical Memorandum 83-2).