Proceedings of the 1990 Winter Simulation Conference
Osman Balci, Randall P. Sadowski, Richard E. Nance (eds.)

FITTING PARAMETRIC MODELS BY CONDITIONAL SIMULATION

Russell C.H. Cheng

School of Mathematics
University of Wales College of Cardiff
Senghennydd Road
Cardiff CF2 4AG, U.K.

ABSTRACT

The Rao-Blackwell theorem is applied to show that the method of control variates can be effected either in the standard way, or by means of an equivalent conditional sampling procedure where the control variates are, in essence, stratified. This alternative method, which we call ***conditional simulation***, is particularly convenient if a parametric model is to be fitted to the simulation response. An application which estimates the saturation point of a single server queue is described.

1. INTRODUCTION

The Rao-Blackwell theorem is a well known result of mathematical statistics. We show that it can be applied to examine the method of control variates. The theorem gives rise to two distinct approaches to how control variates can be utilized. One is the traditional method which is essentially that of linear regression (see for example Venkatraman and Wilson, 1986). The second can be interpreted as the stratified sampling of control variates. We give conditions under which the two methods are equivalent, when they give equal variance reduction.

From the practical viewpoint the second approach of using stratified sampling has a number of features of interest. It allows simulation runs to be carried out conditional on control variates being set equal to selected prescribed values. Arguably this is more in the spirit of what a user would really like to do in a simulation. For example in a single server queue, it may be of interest to examine the queue's behaviour at certain prescribed traffic intensities. Conditional sampling would call for the ***sample*** traffic intensity to be prescribed and then the simulation run to be made conditional on the traffic intensity being set equal to this given value. We call this method of simulation ***conditional simulation.***

A complementary aspect of this approach is to utilize a parametric model to try to define the response of a simulation more precisely. This could be done with the usual method of control variates, but the conditional simulation approach allows the regressor function to be thought of as a fixed known function rather than being random.

In the next Section we describe the underlying theory. In Section 3 we describe some sampling schemes which can be used to control the sampling of traffic intensities, when inter-arrival and service times are gamma distributed. An application to the M/M/1 queue is considered in detail.

2. THEORY

Suppose that a simulation run yields a response Y and a control variate X which we can assume to be related by the equation

$$Y = \tau X + Z \tag{2.1}$$

where $\tau = \text{cov}(X,Y)/\text{var}(X)$; τ will usually be unknown. We shall write μ_X for E(X), $\sigma_X{}^2$ for var(X), and so on, for convenience. Assume also that N independent simulation runs are made yielding the observations (X_i, Y_i) i = 1, 2, ..., N. The objective is to estimate μ_Y, the mean of Y, from these observations.

Much of what follows can be developed without additional assumptions, but for clarity we shall assume initially, if the simulation runs are allowed to become long, that asymptotically X and Y become jointly normally distributed.

Now recall the Rao-Blackwell theorem (see Fraser, 1967 for example):

<u>**Theorem**</u>. Let Y_i, i = 1, 2, ..., N be a random sample from a distribution with density $f_Y(\boldsymbol{\theta})$ and let $\hat{\theta}_1$ estimate θ_1, the first component of $\boldsymbol{\theta}$. Suppose **T** is a sufficient for $\boldsymbol{\theta}$, and let $\tilde{\theta}_1 = E(\hat{\theta}_1 | \mathbf{T})$. Then $\text{var}(\tilde{\theta}_1) \le \text{var}(\hat{\theta}_1)$.

The theorem is usually stated under the additional condition that $\hat{\theta}_1$ is unbiased, when $\tilde{\theta}_1$ is also unbiased; but this is not strictly necessary.

We apply the theorem to (2.1) under the assumption:

(A) X,Y are jointly normal with μ_X, $\sigma_X{}^2$, τ known.

<u>**Theorem 1**</u> Under assumption (A):

$$\text{Var}(\tilde{\mu}_Y) \le \text{Var}(\bar{Y}) \tag{2.2}$$

where

$$\tilde{\mu}_Y = E(\bar{Y} | \mathbf{Z}) = \bar{Y} - \tau(\bar{X} - \mu_X).$$

<u>**Proof**</u> Under assumption (A) the only unknown parameters are μ_Z and $\sigma_Z{}^2$; **Z** is sufficient for these. If we set $\theta_1 = \mu_Y$ and apply the theorem to $\hat{\theta}_1 = \bar{Y}$ we get (2.2) immediately. To verify the last part we have:

$$\begin{aligned} \tilde{\theta}_1 &= E(\hat{\theta}_1 | \mathbf{Z}) \\ &= E_Y(\bar{Y} | \mathbf{Z}) \\ &= E_X(\tau\bar{X} + \bar{Z} | \mathbf{Z}) \\ &= \tau\mu_X + \bar{Z} \\ &= \bar{Y} - \tau(\bar{X} - \mu_X), \end{aligned} \tag{2.3}$$

as required. □

The estimator $\tilde{\theta}_1$ is clearly the conventional control variates estimator.

The condition that τ be known can be relaxed at this stage on noting that if we replace τ by an estimator $\hat{\tau}$ with bias $O_p(N^{-1})$ and variance also $O_p(N^{-1})$, then the asymptotic distribution of $\tilde{\theta}_1$ remains unchanged. Thus, for example, the least squares estimator for τ would do.

There are two points to note about the preceding result. Firstly it works by "integrating out" the variation in the control variate X. Secondly it does this from an "undoctored" random sample. We now show that in situations where the sampling of X can be controlled, we can achieve the same effect by means of stratified sampling.

<u>Theorem 2</u> Let (ξ_i, Y_i), i = 1, 2, ..., N be a sample where the ξ_i's are ***fixed***. Define

$$\hat{\mu}_Y = \bar{Y} = \tau\bar{\xi} + \bar{Z}. \quad (2.4)$$

Then under assumption (A) and if $\bar{\xi} = \mu_X$, the distributions of $\hat{\mu}_Y$ and $\tilde{\mu}_Y$ are identical.

Proof: Under assumption (A), and if $\bar{\xi} = \mu_X$, then $\hat{\mu}_Y$ is identical to (2.3) and the result follows. □

In practice the condition $\bar{\xi} = \mu_X$ can be achieved in at least three separate ways

(M1) Set $\xi_i = \mu_X$ i = 1, 2, ..., N

(M2) Use a fixed set of ξ_i values which in effect make a numerical quadrature of $\int x f_X(x)dx$ so that $\bar{\xi} = \mu_X$.

(M3) The analogue of M2, but carried out by stratified sampling from the distribution of X. In this case care is needed to ensure that $\text{Var}(\bar{\xi}) = o(N^{-1})$ so that the asymptotic distribution of $\hat{\mu}_Y$ is not altered.

In selecting which of the three options to use it is important to take account of how the overall method should be implemented and the effect of assumption (A) not being exactly satisfied. In this respect option M3 is best as stratification of X will not alter the distribution of Y so that $\hat{\mu}_Y$ will be unbiased whether (A) holds or not. In contrast M1 does require the exact independence of X and Z to work without error. However M1 does allow the runs to be pinpointed at the control variate value of most immediate interest. A good compromise would therefore seem to be option M2. A set of ξ_i values can be prescribed allowing conditional simulation at selected (interesting) values of the control variate.

A natural adjunct of this idea is to fit a parametric model to the response variable treating the control variable as, in effect, a deterministic independent variable on which Y depends. Thus we assume

$$Y = \eta(X,\theta) + Z(\theta). \quad (2.5)$$

We make the runs using non-random quadrature of X, and estimate θ using least squares. The advantage of this approach is that it allows for much more specific models to be tried in characterizing the behaviour of Y. An example occurs in the estimation of the traffic intensity at which a queueing system saturates and becomes unstable. We discuss such an example in the next Section.

3. A QUEUEING EXAMPLE

To illustrate the ideas of the previous section we consider the M/M/1 queue with Poisson arrivals, rate λ, and exponential service times with parameter μ. We wish to estimate the average waiting time W in the system (including service) of a customer, and see how it depends on the traffic intensity ρ. The known standard result is that $E(W) = (\mu - \lambda)^{-1}$. We pretend not to know this and examine its estimation by conditional simulation of n customers.

Before describing how each simulation run is to be done it will be necessary to gather together some properties of the gamma distribution.

Let t_i, i = 1, 2, ..., n be the inter-arrival times and s_i, i = 1, 2, ..., n the service times of the n customers. Let

$$T = \sum_{i=1}^{n} t_i \quad , \quad S = \sum_{i=1}^{n} s_i. \quad (3.1)$$

The λT and μS both have the gamma distribution G(n) with pdf

$$f_G(x, n) = \frac{x^{n-1}}{\Gamma(n)} e^{-x} \ , \ x > 0. \quad (3.2)$$

Moreover we have:

Lemma

Let R have the beta distribution Be(n,n) and C the gamma distribution G(2n), with R and C independent. Then

$$\lambda T = C(1-R) \text{ and } \mu S = CR$$

are independent G(n) variates.

<u>Proof</u> This is a standard result of gamma and beta variates (see Aitchison, 1963, or Rao 1973, for example), obtainable by transforming the joint distribution of R and C to that of λT and μS. □

Sampling R (and C) and then forming λT and μS in effect allows the sample traffic intensity

$$\bar{\rho} = \frac{\lambda}{\mu} \frac{R}{(1-R)}$$

to be controlled by controlling R.

Using this result the basic conditional simulation run may be arranged as follows:

(i) Generate R from Be(n,n).
(ii) Generate C from G(2n).
(iii) The sums of the inter-arrival and service times can then be obtained as

$$T = C(1-R)/\lambda \quad \text{and} \quad S = CR/\mu$$

(iv) The individual t_i's and s_i's are then obtained as independent G(1) variates t_i', s_i' i = 1, 2. ..., n which can then be rescaled (see Cheng, 1984) as

$$t_i = \frac{t_i}{\Sigma t_i'} . T \ , \quad s_i = \frac{s_i}{\Sigma s_i'} . S \ i = 1, 2, \ldots, n.$$

This conditional generation of the t's and s's is distributionally exact.

Sampling of R and C can be controlled if they are generated by the inverse probability transform method. If F_R and F_C are the cdf's of R and C respectively then we can generate R and C using

$$R = F_R^{-1}(U) \qquad C = F_C^{-1}(V)$$

where 0 < U, V < 1.

The inverse distribution function F_R^{-1} can be computed using the algorithm given by Cran, Martin and Thomas (1977). However for large n, the distribution is very closely approximated by the normal with mean equal to ½ and variance equal to $[4(2n+1)]^{-1}$. We can then use the Hastings approximation to the inverse of the normal cdf as given by Abramowitz and Stegun (26.2.23, 1965).

For the inverse distribution function F_C^{-1} we solve the equation

$$F_C(c) = V$$

for c using Newton-Raphson iterations using the mode of the density $c_0 = 2n-1$ as starting value (see Devroye, 1986). F_C can be computed using the algorithm AS239 given by Shea (1988).

As only one value of R and C is needed in each run, the added expense of generating these in this way rather than by library routines is negligible.

The above method allows a block of runs to be made with preselected values of R and C using any of the methods M1, M2 or M3 of the previous section.

As an illustration we consider method M1 where R and C are set at their mean values in each run of the block. Table 1 gives the mean and variance estimates from each of three blocks of a hundred runs each, with each run simulating 5000 customers. For comparison the Table also gives the mean and variance of three blocks where R and C are independently generated from their respective population distributions. The variance reduction is greater than that reported by Morgan (1984) using antithetic varietes.

As a second example we consider method M2 with preselected values of R and C. We use $N = (m-1)^2$ combinations of R and C:

$$(R_i, C_j) : R_i = F_R^{-1}(i/m),\ C_j = F_V^{-1}(j/m),\ 1 \le i,\ j \le m-1,$$

Table 2a gives the values of W corresponding to 81 runs where R and C take each of 9 different values (viz where m=10).

The average of the W's can be used to estimate E(W), however we consider instead their use in estimating the saturation point of the queue by fitting a parametric model to the response. As with any procedure of this sort there is the possibility of biasing error occurring through incorrect model selection, and this has to be offset against the benefits of fitting a model with structure. In our example we illustrate by fitting the model (2.5) with X = (R,C) and

$$\eta(R,C;\theta) = \frac{\theta_1(1 - (\theta_2\bar{\rho})^{\theta_3 n})}{\bar{\mu}(1 - \theta_2\bar{\rho})}$$

where $\bar{\rho} = \lambda R/[\mu(1-R)]$, is the sample traffic intensity and $\bar{\mu} = \mu n/(CR)$, is the sample service rate, and $\theta = (\theta_1, \theta_2, \theta_3)$ is the vector of parameters to be estimated. The form of η has the correct general behaviour in that, if $\theta_2\bar{\rho} < 1$, η remains finite as $n \to \infty$, but $\eta \to \infty$ as $n \to \infty$ when $\theta_2\bar{\rho} \ge 1$. Thus in effect θ_2 estimates the saturation level of the queue. Functionally η does not have the correct form for the case when $\theta_2\bar{\rho} > 1$ but the inclusion of the additional parameters θ_1 and θ_3 builds, hopefully, some tolerance into the model. A more sophisticated model would perhaps incorporate different functional forms for η depending on whether $\theta_2\bar{\rho}$ is less than or greater than 1.

A least squares fit gave

$$\hat{\theta} = (.878,\ 1.027,\ 1.100)$$

and the fitted function, $\hat{\eta}$, is tabulated in Table 2b together with the residuals in Table 2c. The value of $\hat{\theta}$ estimates the equilibrium expected value of W as

$$\hat{E}(W) = \frac{.878}{\mu - 1.027\lambda}$$

compared with the known exact value of $E(W) = (\mu - \lambda)^{-1}$.

In conclusion we remark that the above method offers clear advantages over straight replication where N independent but identical runs are made. By conditioning on control variates it allows these to be treated like deterministic regressor variables so that simulation runs can be done at prescribed control variate values. This extends naturally to the use of parametric models for the mean of the response of interest. This yields additional variance reduction as the fitted regression will have less variance than individual observations, and moreover allows for a more insightful interpretation of the response behaviour through the parametric model.

Table 1. The Sample Means and Variances of 100 observations from the simulation of an M/M/1 queue starting empty, $\lambda=0.8$, $\mu=1$. Each observation is the average waiting time of 5000 customers. Three sets of results are given, each for the case when the runs are independent, and for the case when method M1 is used.

Independent Runs		*Runs using Method M1*	
Mean	Variance	Mean	Variance
5.02	.337	4.93	.135
5.03	.489	4.94	.182
4.92	.329	4.86	.160

Table 2. Average Waiting Times of 5000 Customers in an M/M/1 at 81 Different Combinations of Values of the Control Variates R and C. λ=0.8, μ=1.

(a) Observed Average Waiting Times

R \ C	9872	9916	9947	9974	10000	10025	10052	10084	10028
0.494	4.50	4.17	3.93	4.50	4.13	4.37	4.26	4.44	4.21
0.496	4.80	5.03	4.42	4.04	4.83	4.59	5.03	4.33	5.66
0.497	5.33	4.26	4.45	3.97	5.70	5.35	5.02	4.76	5.19
0.499	4.80	4.93	4.57	4.99	5.20	4.57	4.16	4.78	4.81
0.500	4.35	4.81	4.85	5.48	4.64	4.63	4.47	4.14	4.79
0.501	5.79	5.21	4.49	5.28	5.82	5.04	5.17	4.78	4.91
0.503	4.49	5.31	5.78	5.00	4.71	4.85	5.26	5.25	5.10
0.504	4.73	6.17	4.79	5.08	4.79	5.31	4.75	5.21	4.96
0.506	4.83	5.74	6.01	6.26	8.74	5.03	5.27	5.82	6.10

(b) Fitted Model

R \ C	9872	9916	9947	9974	10000	10025	10052	10084	10028
0.494	4.31	4.32	4.34	4.35	4.36	4.37	4.38	4.40	4.42
0.496	4.48	4.50	4.52	4.53	4.54	4.55	4.57	4.58	4.50
0.497	4.62	4.64	4.66	4.67	4.68	4.69	4.71	4.72	4.74
0.499	4.75	4.77	4.78	4.80	4.81	4.82	4.83	4.85	4.87
0.500	4.87	4.89	4.91	4.92	4.93	4.95	4.96	4.98	4.00
0.501	5.00	5.02	5.04	5.05	5.06	5.08	5.09	5.11	5.13
0.503	5.15	5.17	5.19	5.20	5.21	5.23	5.24	5.26	5.28
0.504	5.33	5.35	5.37	5.38	5.40	5.41	5.43	5.44	5.47
0.506	5.60	5.63	5.65	5.66	5.68	5.69	5.71	5.73	5.75

(c) Residuals

R \ C	9872	9916	9947	9974	10000	10025	10052	10084	10028
0.494	0.19	-0.15	-0.41	0.15	-0.23	0.00	-0.13	0.05	-0.21
0.496	0.32	0.53	-0.09	-0.49	0.29	0.04	0.46	-0.25	1.06
0.497	0.71	-0.39	-0.20	-0.80	0.02	0.65	0.31	0.04	0.45
0.499	0.05	0.16	-0.22	0.19	0.39	-0.25	-0.67	-0.07	-0.06
0.500	-0.52	-0.08	-0.06	0.56	-0.29	-0.32	-0.49	-0.84	-0.20
0.501	0.79	0.18	-0.55	0.23	0.76	-0.03	0.08	-0.33	-0.22
0.503	-0.66	0.14	0.60	-0.20	-0.50	-0.37	0.02	-0.00	-0.18
0.504	-0.60	0.82	-0.58	-0.30	-0.61	-0.10	-0.67	-0.23	-0.50
0.506	-0.77	0.11	0.36	0.60	0.06	-0.66	-0.43	0.10	0.35

REFERENCES

Abramowitz, M. and I.E. Stegun (1970), *Handbook of Mathematical Functions*, Dover Publications, New York, NY.

Aitchison, J. (1963), "Inverse Distributions and Independent Distributed Products of Random Variables," *Biometrika 50*, 505-508.

Cheng, R.C.H. (1984), "Generation of Inverse Gaussian Variates with Given Sample Mean and Dispersion," *Journal of Applied Statistics 33*, 309-316.

Cran, G.W., K.J. Martin, and G.E. Thomas (1977), "Algorithm AS109: Inverse of the Incomplete Beta Function Ratio," *Journal of Applied Statistics 26*, 111-114.

Devroye, L. (1986), *Non-Uniform Variate Generation*, Springer-Verlag, New York, NY.

Fraser, T.S. (1967), *Mathematical Statistics*, Academic Press, New York, NY.

Morgan, B.J.T. (1984), *Elements of Simulation*, Chapman and Hall, London.

Rao, C.R. (1973), *Linear Statistical Inference and its Applications*, Wiley, New York, NY.

Shea, B.L. (1988), "Algorithm AS239 Chi-Squared and Incomplete Gamma Integral," *Journal of Applied Statistics 37*, 466-473.

Venkatraman, S. and J.R. Wilson (1986), "The Efficiency of Control Variates in Multi-Response Simulation," *Operations Research Letters 5*, 37-42.