Proceedings of the 1987 Winter Simulation Conference
A. Thesen, H. Grant, W. David Kelton (eds.)

SMARTER CONTROL VARIABLES: REGRESSION-ADJUSTED LINEAR AND NONLINEAR CONTROLS

Peter A.W. Lewis
Richard Ressler
R. Kevin Wood
Naval Postgraduate School
Monterey, CA 93943, U.S.A.

ABSTRACT

Nonlinear regression-adjusted control variables are investigated for improving variance reduction in statistical and system simulations. Simple control variables are transformed using linear and nonlinear transformations, and parameters of these transformations are selected using linear or nonlinear least squares regression. As an example, piecewise power-transformed variables are used in the estimation of the mean for the two variable Anderson-Darling goodness-of-fit statistic W_2^2. Substantial variance reduction over straightforward controls is obtained. These parametric transformations are compared against optimal, additive, nonparametric transformations from ACE and are shown to be nearly optimal.

1. PRELIMINARIES

This paper investigates the use of possibly nonlinear. regression-adjusted control variates for variance reduction in statistical and system simulation.

Let C be a vector of control variables which are correlated to a statistic of interest Y, and assume that C has known mean vector E(C). The standard method of obtaining a controlled statistic Y' to estimate $E(Y)$, and which has less variance than Y, is via the linear, additive combination

$$Y' = Y - \beta^{\mathrm{T}}(\mathrm{C} - \mathrm{E}(\mathrm{C})). \tag{1}$$

The vector β is a vector of unconstrained constants chosen to minimize the variance of Y'. Note that some components of C may be known power transformations of other components, so that polynomial control schemes are included in formulation (1). Explicit expressions for the components of β which minimize the variance of Y' can be found in terms of the second order moments of Y and C, and with these parameters, Y' is an unbiased estimate of $E(Y)$.

This paper generalizes (1) by letting

$$Y' = Y - C', \tag{2}$$

where C' is any mean-zero linear or nonlinear parametric function of the components of C, i.e.. $C'=f(\mathrm{C};\beta)-\mathrm{E}(f(\mathrm{C};\beta))$. For example, C' might involve additive or multiplicative combinations of power transformations of the components of the original control vector C. Optimal or near-optimal values of the unknown parameters of those transformations. analogous to β in (1), are obtained by minimizing the variance of Y', but the results are not explicit functions of the joint and higher moments between Y and the set of control variables. Before going on to the details of the general case (2) of regression-adjusted controls, we review simple linear controls.

Consider the case of a single, additive, linear control so that $C'=\beta C-\beta E(C)$. Then,

$$Y' = Y - \beta(C - \mathrm{E}(C)),$$

and β is chosen to minimize var(Y'). This variance is minimized when β is proportional to the correlation between C and Y; the greater the correlation, the greater the effectiveness of the control in obtaining variance reduction. Assuming var(Y)= var(C), the result follows from:

$$\begin{aligned}\mathrm{var}(Y') &= \mathrm{var}(Y) + \beta^2\mathrm{var}(C) - 2\beta\mathrm{cov}(Y,C)\\ &= \mathrm{var}(Y)(1 + \beta^2 - 2\beta\rho(Y,C)).\end{aligned}$$

Differentiating with respect to β and setting the resulting expression equal to zero yields the optimal value for β:

$$\beta = \rho(Y,C),$$

where

$$\frac{\mathrm{var}(Y')}{\mathrm{var}(Y)} = 1 - \rho(Y,C)^2. \tag{3}$$

In particular,

$$100\left(\frac{\mathrm{var}\,Y - \mathrm{var}\,Y'}{\mathrm{var}\,Y}\right) = 100\left(1 - \frac{\mathrm{var}\,Y'}{\mathrm{var}\,Y}\right) \tag{4}$$

measures the percent variance reduction resulting from the control. Without the assumption of equal variances, we have

$$\beta = \rho(Y,C)\sigma_Y/\sigma_C,$$

while (3) still holds. Thus, if $\rho(Y,C)$, σ_Y and σ_C are known, $\rho(Y,C)$ is a direct measure of the variance reduction which can be obtained with a single regression-adjusted control.

Now, consider the more general case of multiple, possibly nonlinear, control variables. Using (2). we obtain

$$\frac{\mathrm{var}(Y')}{\mathrm{var}(Y)} = 1 + \frac{\mathrm{var}(C')}{\mathrm{var}(Y)} - 2(\sigma_{C'}/\sigma_Y)\rho(Y,C') \quad (5)$$

$$= 1 + k^2 - 2k\rho(Y,C')$$

and

$$1 - \frac{\mathrm{var}(Y')}{\mathrm{var}(Y)} = 2k\rho(Y,C') - k^2, \quad (6)$$

where k is positive valued. While this last equation is simple in form, both $\rho(Y,C')$ and $k = \sigma_{C'}/\sigma_Y$ are functions of the parameters in C'. Thus, it is not true that in order to maximize the variance reduction with respect to the parameters of the control function, one need only maximize the absolute value of the correlation between Y and C'.

When C' is a linear, additive function of the components of C, as in (1), $\rho(Y,C')$ is a quadratic function of the parameters β whose optimal values are a function of the correlation matrix of (Y,C), i.e., the joint and higher moments between Y and the set of control variables. In fact, explicit expressions for the optimal values of β are known (Rubenstein and Marcus, 1985).

For two independent linear controls with known correlations with Y it follows from (5) that with the optimal values of β,

$$\frac{\mathrm{var}(Y')}{\mathrm{var}(Y)} = 1 - \rho(Y,C_1)^2 - \rho(Y,C_2)^2. \quad (7)$$

Choosing control variables with maximum correlations with Y will, in this case, still maximize the reduction in variance. In general when the controls are not independent, and $\sigma_{C'}/\sigma_{Y'} \neq \rho(Y,C')$, $\rho(Y,C')$ does not yield an exact measure of variance reduction. Note too that in the general case (2), the allowable range of parameters in the function C' of the components of C may be constrained by the requirement that $E(C')$ must be known, exactly or approximately, and must be finite.

2. THE SAMPLE ANALOG TO THE VARIANCE REDUCTION FORMULA

In practice, one has no theoretical information about properties of Y and C, but one has a simulation sample of size m of independent replications $\{Y_i,\mathrm{C}_i\}$ from which to estimate $E(Y)$. Regardless of whether the sample is large or small, one wants to minimize the sample variance of Y'. Minimizing the sample variance involves, after subtracting $\overline{Y}$ from both sides of (2), minimizing

$$\frac{\sum(Y_i'-\overline{Y})^2}{m} = \frac{\sum(Y_i-\overline{Y}-C_i')^2}{m} \quad (8)$$

$$= \frac{\sum(Y_i-\overline{Y})^2}{m} + \frac{\sum C_i'^2}{m} - \frac{2\sum(Y_i-\overline{Y})C_i'}{m} \quad (9)$$

The left-hand side of (8) is the quantity to be minimized since $\mathrm{E}(\overline{Y})=\mathrm{E}(\overline{Y}')=\mathrm{E}(Y)$ if $E(C')$ is known. Equation (8) shows that this quantity is equal to the residual sum of squares of the least squares regression of $Y_i-\overline{Y}$ on C'. Equation (9) involves, in its first term, the total sum of squares, which estimates the variance of Y; in its second term the sample variance of the zero mean C'; and in the last term the sample covariance of Y and C'. Rearranging terms in (9), we have

$$\frac{\sum(Y_i-\overline{Y})^2}{m} - \frac{\sum(Y_i'-\overline{Y})^2}{m} = \frac{2\sum(Y_i-\overline{Y})C_i'}{m} - \frac{\sum C_i'^2}{m}$$

or

$$\frac{\sum(Y_i-\overline{Y})^2-\sum(Y_i'-\overline{Y})^2}{\sum(Y_i-\overline{Y})^2} = \frac{2\sum(Y_i-\overline{Y})C_i'}{\sum(Y_i-\overline{Y})^2} - \frac{\sum C_i'^2}{\sum(Y_i-\overline{Y})^2}. \quad (10)$$

The left-hand side of (10) is the usual R^2 regression measure and the equation may be rewritten as

$$R^2 = 2r(Y,C')\frac{S_{C'}}{S_Y} - \frac{S_{C'}^2}{S_Y^2} = 2\,\hat{k}r(Y,C') - \hat{k}^2. \quad (11)$$

As the sample analog to (6), (11) indicates that maximizing R^2 through nonlinear least squares regression methods is equivalent to maximizing variance reduction when the optimal parameters are unknown.

Thus, for multivariate C, this maximization can be accomplished through multiple least squares regression of $Y'-Y$ on C'. With linear controls, linear least squares regression will provide a global minimum for the residual sum of squares, in turn maximizing the variance reduction for the sample. Using the regression-derived β in the control equation (2) maximizes the correlation between Y and C' for the particular sample. When the control function is nonlinear, nonlinear least squares regression will not necessarily determine parameter values which globally minimize the residual sum of squares since nonconvexity of the control function may create suboptimal local minima. With a control function $C'=f(\mathrm{C};\beta)-E(f(\mathrm{C};\beta))$ that is nonconvex, the choice of initial values for the parameters β in the nonlinear regression may significantly affect the amount of variance reduction obtained. In both the convex and nonconvex cases, bounds on the values of the parameters may be necessary to ensure valid transformations. One must also be careful that while multiple regression can be computationally useful, the distribution theory behind multiple regression, which assumes fixed independent variables, does not apply. Consequently, confidence intervals on parameter estimates cannot be determined.

3. ESTIMATING CONTROL PARAMETERS:

In the case of the single linear control with a fixed parameter β, Y' is an unbiased estimator of Y since $\mathrm{E}(C-\mathrm{E}(C))=0$. This is also true for multiple linear controls when β is fixed. When $\rho(Y,C')$, σ_Y and $\sigma_{C'}$ are not known, they must be estimated from data. Such estimates can then be used to compute β for a single control. For multiple controls, the optimal β can be estimated using additive, least squares, multiple regression on a data sample. If the data sample from which the estimates are derived is the one to be controlled, the estimates of the control parameters may be biased and the effectiveness of the control may be reduced when used with other samples. Using a data sample other than the one to be controlled, such as a small test sample, will eliminate the bias that arises from the lack of independence between the control parameters and the sample. However, the problem of small sample bias may then arise. For any sized sample, there is the additional problem of estimating the variance of the regression-adjusted estimate of $E(Y)$.

4. PIECEWISE LINEAR TRANSFORMATIONS OF CONTROLS

Statistics are often nonlinear functions of the random variables from which they are derived. Therefore one might expect some nonlinear controls to have a higher correlation with Y than linear controls, and therefore, roughly, to be able to better "control" than the linear controls. One type of nonlinear control can be formed from an initial guess at a viable control by the use of indicator functions and "cutpoints" to form piecewise linear transformations of controls. For example a control variable C is split into two control variables about a cutpoint δ as follows:

$$C_1 = \begin{cases} C \text{ if } X \leqslant \delta \\ 0 \text{ otherwise,} \end{cases} \quad C_2 = \begin{cases} C \text{ if } X > \delta \\ 0 \text{ otherwise.} \end{cases} \qquad (12)$$

By judicious choice of the cutpoint or perhaps multiple cutpoints, least squares multiple regression can achieve a better fit without the use of additional original variables. Of course, care must then be taken in determining the form of the control function to ensure it has mean zero. Note also that the regression is still linear if δ is given, but it is nonlinear otherwise. Then, finding an optimal δ becomes, in general, a nonconvex, nonlinear, mathematical programming problem.

5. POWER TRANSFORMATIONS OF CONTROLS

Power transformations of controls, in addition to piecewise transformations of controls, introduce nonlinearity into the controlled estimate of $E(Y)$. The power transformation used initially in this study is of the form $(C^p-1)/p$, for $p>-1$. This scaled power transformation is equal to $\ln C$ when $p=0$. Using, for example, the single control variable C, the resulting control function is

$$C' = \beta\left\{\frac{C^p-1}{p} - \frac{\mathrm{E}(C^p-1)}{p}\right\}$$

which has two parameters, p and β. Of course, piecewise transformations of power transformations are also possible, and it is this combination of nonlinear controls which is the main thrust of this paper. One hopes to come close to the maximum theoretical variance reduction which could be obtained.

6. THE ACE PROGRAM

The ACE (Alternating Conditional Expectation) program (Breiman and Freidman, 1985) provides a method for estimating the minimum variance obtainable by regressing a variable Y on an additive combination of arbitrary transformations of another set of variables such as $\mathbf{C}$. It uses an iterative algorithm to do this. The procedure is nonparametric, with the transformations selected solely on the basis of the data sample. Minimal assumptions about the distribution of the sample or allowable transformations enable ACE to produce an estimate of the minimum mean square error between the transformed Y variable and the sum of the transformed components of $\mathbf{C}$. When $\mathbf{C}$ has only one component, this is equivalent to maximizing the correlation between a transformed Y and a transformed $\mathbf{C}$.

Unfortunately, the transformations ACE uses cannot be used to develop control variables as they are non-parametric and the true means of the transformed variables cannot be determined. However, one can use the minimum mean square error from ACE to obtain an upper bound on the variance reduction that can be achieved between Y and C' in a parametric control function such as (2). Thus, ACE may be used to gauge the effectiveness of any control function using a fixed set of control variables.

7. AN EXAMPLE

Estimating the mean of the Anderson-Darling goodness-of-fit statistic (Anderson and Darling 1952) provides a good example of the benefits of piecewise controls and power transformations. The statistic W_n^2 can be determined as a function of n independent unit exponential random variables E_k (Lewis and Orav, 1987). The independence of these random variables makes them ideal for controlling W_n^2. The case $n=2$ is presented here, for which (7) holds with $C_1=E_1$ and $C_2=E_2$.

Five different control functions were evaluated using a single sample of 500 pairs of unit exponentials and their associated W_2^2 values. The following control functions were compared:

$$C' = \beta_1(E_1-1) + \beta_2(E_2-1); \tag{13a}$$

$$C' = \beta_1(E_1-1)+\beta_2(E_2-1)+\beta_3(E_1^2-2)+\beta_4(E_2^2-2); \tag{13b}$$

$$C' = \sum_{j=1}^{2}\beta_j\left[\frac{E_j^{p_j}-1}{p_j} - \mathrm{E}\left(\frac{E_j^{p_j}-1}{p_j}\right)\right]; \tag{13c}$$

$$C' = \sum_{j=1}^{2}\sum_{k=1}^{2}\beta_{jk}\left[\frac{E_{jk}^{p_{jk}}-1}{p_{jk}} - \mathrm{E}\left(\frac{E_{jk}^{p_{jk}}-1}{p_{jk}}\right)\right], \tag{13d}$$

where

$$E_{j1} = \begin{cases} E_j \text{ if } E_j \leqslant \delta_j \\ 0 \text{ otherwise} \end{cases} \quad E_{j2} = \begin{cases} 0 \text{ if } E_j \leqslant \delta_j \\ E_j \text{ otherwise} \end{cases} \quad j=1,2;$$

$$C' = \sum_{j=1}^{2}\sum_{k=1}^{3}\beta_{jk}\left[\frac{E_{jk}^{p_{jk}}-1}{p_{jk}} - \mathrm{E}\left(\frac{E_{jk}^{p_{jk}}-1}{p_{jk}}\right)\right], \tag{13e}$$

where

$$E_{j1} = \begin{cases} E_j \text{ if } E_j \leqslant \delta_{j1} \\ 0 \text{ otherwise} \end{cases} \quad E_{j2} = \begin{cases} E_j \text{ if } \delta_{j1} < E_j \leqslant \delta_{j2} \\ 0 \text{ otherwise} \end{cases}$$

$$E_{j3} = \begin{cases} E_j \text{ if } E_j > \delta_{j2} \\ 0 \text{ otherwise} \end{cases} \quad j=1,2.$$

The experimental, APL-based GRAFSTAT, from IBM Research, was used for all of the computing. Controls (13a) and (13b) were developed using straightforward least squares regression. Controls (13c), (13d), and (13e) were developed using the nonlinear regression segment of GRAFSTAT. For nonlinear regression, GRAFSTAT uses a form of the Marquadt algorithm (Marquadt, 1963) which allows bounds to be placed on the parameters. For controls (13c), (13d), and (13e), which have powers as parameters, lower bounds of $-.99$ were necessary on the power parameters p_{jk} since expected values of the truncated exponential variables (involving the gamma function) are not defined for $p_{jk} \leqslant -1$. A reasonable upper bound on each p_{jk} was found useful in speeding convergence.

The cutpoints for (13d) and (13e) were fixed at quantiles such as the .5 or .33 and .66 quantiles and were not included as parameters in the optimization. Although the two cutpoints employed on (13e) could be used to divide the plane into nine regions, the six marginal variables were used for control. The control function thus contained two independent sets of three controls with simple distributions, versus nine controls whose expected values involve multivariate distributions.

As expected, the simplest control was the least effective. Control (13a) achieved an R^2 of .2265, which is hardly worthwhile. Control (13b), which is a "standard" control in that the powers are fixed, gave an R^2 of .5629. The most complex control, (13e), had the highest R^2 at .8354. The R^2 value derived by ACE was .8560 showing that control (13e) is nearly optimal for the control variables used.

8. SUMMARY AND CONCLUSIONS

This study demonstrates the potential effectiveness of nonlinear regression-adjusted controls in reducing variance in simulations. Various piecewise linear and power transformations were shown to be useful in developing control functions. Other topics to be studied include:

(a) Finding controls for the variance, percentiles and quantiles of W_2^2;

(b) Finding controls for W_n^2 for $n>2$, perhaps using measures of influence or leverage to reduce the size of the control function;

(c) Using other transformations such as

(1) $(e^{\gamma X}-1)/\gamma$,

(2) $(X^p e^{\gamma X}-1)/p\gamma$, or

(3) $(e^{(X^p-1)/p}-1)/\gamma$;

(d) Using similar controls for gamma family statistics such as those encountered in queuing problems.

(e) Investigating problems of estimating the variance of the variance-reduced estimate of $E(Y)$.

ACKNOWLEDGEMENTS

The research of P.A.W. Lewis was supported at the Naval Postgraduate School under Office of Naval Research Grant RR014-05-01. We are indebted to Dr. P.D. Welch for permission to use GRAFSTAT under a test site agreement with IBM Research.

REFERENCES

Anderson, T.W. and Darling, D.A. (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic process. *Annals of Mathematical Statistics*, **23**, 193-212.

Breiman, L. and Freidman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, **80**, 580-619.

Lewis, P.A.W. and Orav, E.J. (1987). *Simulation Methodology for Statisticians, Engineers, and Operations Analysts*, to be published.

Marquadt, D.W. (1963), An algorithm for least squares estimation of nonlinear parameters, *SIAM Journal*, **11**, 431-441.

Rubenstein, R.Y. and Marcus, R. (1985), Efficiency of multivariate control variates in Monte Carlo simulation, *Operations Research*, **33**, 661-667.

AUTHORS BIOGRAPHIES

PETER A.W. LEWIS is Professor of Statistics and Operations Research at the Naval Postgraduate School in Monterey, California. He has the A.B., B.S., and M.S. degrees in Electronic Engineering from the School of Engineering at Columbia University and a Ph.D. in Statistics from the University of London. He is the co-author, with D.R. Cox, of the book *The Statistical Analysis of Series of Events* and the author, with E.J. Orav, of the forthcoming book *Simulation Methodology for Statisticians, Operations Analysts and Engineers.*

RICHARD L. RESSLER is a Ph.D. student in the Operations Research Department at the Naval Postgraduate School in Monterey, California. He received a B.A. in biochemistry from the University of Pennsylvania in 1978. He is a student member of O.R.S.A..

R. KEVIN WOOD is Associate Professor of Operations Research at the Naval Postgraduate School in Monterey, California. He has B.S. degrees in Mathematics and Electrical Engineering from University of Portland, an M.S. degree in Operations Research from Columbia University, and a Ph.D. degree in Operations Research from University of California, Berkeley. His research areas include network and system reliability and mathematical programming.