CONTROL VARIATE REMEDIES

BARRY L. NELSON

Ohio State University, Columbus, Ohio

(Received August 1988; revision received June 1989; accepted October 1989)

Other than common random numbers, *control variates* is the most promising variance reduction technique in terms of its potential for widespread use: Control variates is applicable in single or multiple response simulation, it does not require altering the simulation run in any way, and any stochastic simulation contains potential control variates. A rich theory of control variates has been developed in recent years. Most of this theory assumes a specific probabilistic structure for the simulation output process, usually joint normality of the response and the control variates. When these assumptions are not satisfied, desirable properties of the estimator, such as unbiasedness, may be lost. A number of remedies for violations of the assumptions have been proposed, including jackknifing and splitting. However, there has been no systematic analytical and empirical evaluation of these remedies. This paper presents such an evaluation, including evaluation of the small-sample statistical properties of the proposed remedies.

Variance reduction techniques (VRTs) are experimental design and analysis techniques used to increase the precision of sampling-based point estimators without a corresponding increase in sampling effort. In this paper, *sampling* means a computer simulation experiment. We only consider one VRT, called *control variates* (CVs), and a specific form of control variates, the *linear* CV. Comprehensive surveys of variance reduction in computer simulation include Wilson (1984) and Nelson (1987a), and more general characterizations of control variates are given by Nelson (1987b) and Glynn and Whitt (1989). For convenience, we use the term *CV* to mean the linear-control variate VRT, and use the term *control* to refer to the auxiliary random variables on which the technique is based.

CVs have the potential for widespread, even automated, use in general stochastic simulation experiments. Using CVs is feasible in *any* stochastic simulation, and applying CVs to estimate one parameter does not conflict with estimating other parameters. Also, there is a rich theory of CVs for a variety of problems, including estimating a univariate mean (Lavenberg and Welch 1981), a multivariate mean (Rubinstein and Markus 1985, Venkatraman and Wilson 1986), and linear metamodels (Nozari, Arnold and Pedgen 1984, Porta Nova and Wilson 1986, Tew and Wilson 1989). Perhaps most importantly, the software required to employ CVs—a simulation language that collects or stores simulation outputs and a least-squares regression package—is readily available.

There are, however, at least three outstanding problems that restrict widespread use of CVs: The problem of selecting effective controls from among the many that are (usually) available, the absence of methods for applying CVs in steady-state simulations with single-replication experiment designs, and the lack of theory or guidance for addressing violations of the standard assumptions on which the theory of CVs is based. The problem of selecting controls has been addressed by Bauer (1987) and Añonuevo and Nelson (1988). Batching and regenerative output analysis methods have been proposed to solve the steady-state simulation problem (e.g., Lavenberg, Moeller and Sauer 1979, Wilson and Pritsker 1984a, b and Nelson 1989). This paper concentrates on violations of the standard assumptions.

The theory of CVs assumes a particular joint distribution for the response variable of interest and the controls, usually multivariate normality. Several remedies for violation of this assumption have been proposed, including jackknifing and splitting. We also consider batching, using known correlation structure, and bootstrapping. These remedies are general purpose, and require little or no special knowledge beyond what is needed to use CVs. We do not consider specialized transformations that might be ideal for certain classes of problems.

There are two fundamental questions regarding any CV remedy: How well does it work when it is needed, and how much does it hurt when it is not needed? In addressing the first question, small-sample properties

Subject classifications: Simulation, efficiency: variance reduction. Simulation, statistical analysis: point and interval estimation.

0030-364X/90/3806-0974 $01.25

of the linear CV and remedies are important, since, asymptotically, violations of the assumptions do not matter. The second question is not always considered, but is equally important, because the standard assumptions can seldom be verified in practice. We should also consider the additional computational burden, if any, imposed by the remedies.

In the next section we define the *crude* experiment, which is where efforts at variance reduction begin. The linear CV and the proposed remedies are introduced and examined next. Finally, we summarize the results of an extensive simulation study of the estimators. All derivations and proofs are given in the appendices or in Nelson (1988).

1. CRUDE EXPERIMENT

We consider estimating a real, scalar parameter θ that is the expected value of an observable random variable Y, denoted $\theta = \mathrm{E}[Y]$. Estimation of multivariate means or metamodels is beyond the scope of this paper, although the estimators we consider can be applied one-at-a-time to estimate a multivariate mean. Let $\sigma_Y^2 = \mathrm{Var}[Y]$.

In the crude experiment we observe $Y_1, Y_2, \ldots, Y_n$, independent and identically distributed (i.i.d.) copies of Y; it may be convenient to think of Y_i as the output from the ith replication in a simulation experiment. The point estimator is the sample mean

$$\hat{\theta}_C = n^{-1} \sum_{i=1}^{n} Y_i$$

and an estimator of $\mathrm{Var}[\hat{\theta}_C] = \sigma_Y^2/n$ is

$$S_C^2 = (n(n-1))^{-1} \sum_{i=1}^{n} (Y_i - \hat{\theta}_C)^2.$$

Both estimators are unbiased.

If Y is normally distributed, then a $(1-\alpha)100\%$ confidence interval for θ is $\hat{\theta}_C \pm H_C$, where $H_C = t_{\alpha/2}(n-1)S_C$, $t_{\alpha/2}(n-1)$ is the $1-\alpha/2$ quantile of the t distribution with $n-1$ degrees of freedom, and $0 < \alpha < 1$. If Y is not normally distributed, then the confidence level is only approximately the nominal $1-\alpha$, but the approximation improves as n increases.

Variance reduction, as the name implies, refers to reducing point-estimator variance. The $\mathrm{Var}[\hat{\theta}_C]$ is the standard against which CV point estimators are compared. However, point-estimator variance is not the only important performance measure. For instance, the CV estimator and some of the remedies may be biased. Let $\hat{\theta}$ be a point estimator of θ. Then its bias is $\mathrm{Bias}[\hat{\theta}] = \mathrm{E}[\hat{\theta} - \theta]$. A summary measure that combines both variance and bias is the mean squared error, $\mathrm{MSE}[\hat{\theta}] = \mathrm{Var}[\hat{\theta}] + \mathrm{Bias}[\hat{\theta}]^2$.

One of the effects of bias is degradation of the probability that the confidence interval covers θ. If the interval is of the form $\hat{\theta} \pm H$, then the probability of coverage is $\Pr\{|\hat{\theta} - \theta| \leqslant H\}$. The probability of coverage may also depend on properties of the estimator of $\mathrm{Var}[\hat{\theta}]$. Coverage greater than the nominal level can always be achieved by making the interval excessively long, so it is important to compare $\mathrm{E}[H]$, the expected halfwidth of the interval, for alternative estimators. These performance measures and others will be used to evaluate the linear CV and the remedies.

2. LINEAR CONTROL VARIATES

Assume that, in addition to the response of interest Y_i, we can also observe a $q \times 1$ vector $\mathbf{C}_i = (C_{i1}, C_{i2}, \ldots, C_{iq})'$, whose mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_q)'$ is known. The $\{\mathbf{C}_i\}$, called the controls, are also i.i.d., but Y_i and $\mathbf{C}_i$ are (hopefully) dependent. Let $\mathbf{Z}_i = [Y_i, \mathbf{C}_i']'$, $i = 1, 2, \ldots, n$ be the $(q+1) \times 1$ vector of the response and controls from the ith replication. Then we assume that the $\{\mathbf{Z}_i\}$ are i.i.d., and let

$$\mathrm{Var}[\mathbf{Z}_i] = \boldsymbol{\Sigma}_{ZZ} = \begin{pmatrix} \sigma_Y^2 & \boldsymbol{\sigma}_{YC} \\ \boldsymbol{\sigma}_{CY} & \boldsymbol{\Sigma}_{CC} \end{pmatrix}$$

where $\boldsymbol{\sigma}_{CY}$ is the $q \times 1$ vector whose jth element is $\mathrm{Cov}[C_{ij}, Y_i]$ and $\boldsymbol{\Sigma}_{CC} = \mathrm{Var}[\mathbf{C}_i]$. All variance-covariance matrices are assumed to be positive definite, and the covariance structure above is assumed to hold throughout the paper. We use the convention that vectors and matrices are indicated by boldface type, and the order of random variable subscripts specifies the dimension of the vector or matrix. Thus, $\boldsymbol{\sigma}_{YC}$ is a $1 \times q$ vector, while $\boldsymbol{\sigma}_{CY}$ is a $q \times 1$ vector. Otherwise, vectors are column vectors.

The type of simulation output process assumed here typically arises in terminating or finite horizon simulations. Examples include: 1) the response Y is the time to complete a stochastic activity network and the control $\mathbf{C}$ is the time to complete selected paths; 2) the response Y is the number of customers served per day in a bank where balking occurs and the control $\mathbf{C}$ is the number of customers that arrive; 3) the response Y is the total cost to operate an inventory system over a finite horizon and the control $\mathbf{C}$ is the total demand over that time horizon; and 4) the response Y is an indicator variable that equals 1 if a new test statistic exceeds a given critical value and $\mathbf{C}$ is the indicator for a second statistic with known power. A queueing example is given in Section 8. Such output processes may also arise in steady-state or infinite horizon simulations when the experiment design specifies replications.

The linear CV point estimator, $\hat{\theta}_L$, is

$$\hat{\theta}_L = \hat{\theta}_C - (\bar{\mathbf{C}} - \boldsymbol{\mu})'\hat{\boldsymbol{\beta}}$$

where $\bar{\mathbf{C}} = n^{-1} \sum_{i=1}^{n} \mathbf{C}_i$, $\hat{\boldsymbol{\beta}} = \mathbf{S}_{\mathrm{CC}}^{-1}\mathbf{S}_{\mathrm{C}Y}$, $\mathbf{C}$ is the $n \times q$ matrix whose ith row is $\mathbf{C}_i'$, $\mathbf{S}_{\mathrm{CC}} = (n-1)^{-1}(\mathbf{C}'\mathbf{C} - n\bar{\mathbf{C}}\bar{\mathbf{C}}')$, $\mathbf{S}_{\mathrm{C}Y} = (n-1)^{-1}(\mathbf{C}'\mathbf{Y} - n\bar{\mathbf{C}}\bar{Y})$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. An estimator of $\mathrm{Var}[\hat{\theta}_L]$ is

$$S_L^2 = S^2(n^{-1} + (n-1)^{-1}(\bar{\mathbf{C}} - \boldsymbol{\mu})'\mathbf{S}_{\mathrm{CC}}^{-1}(\bar{\mathbf{C}} - \boldsymbol{\mu}))$$

where

$$S^2 = (n - q - 1)^{-1} \sum_{i=1}^{n} (Y_i - \hat{\theta}_L - (\mathbf{C}_i - \boldsymbol{\mu})'\hat{\boldsymbol{\beta}})^2.$$

The associated $(1 - \alpha)100\%$ confidence interval is $\hat{\theta}_L \pm H_L$ where $H_L = t_{\alpha/2}(n - q - 1)S_L$.

For the purposes of this paper, the following theorem is the fundamental theorem of CV estimators.

Theorem 1. (Lavenberg and Welch 1981, Appendix A) *Suppose that* $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ *are i.i.d.* $q + 1$*-variate normal vectors with mean vector* $(\theta, \boldsymbol{\mu}')'$ *and variance* $\boldsymbol{\Sigma}_{\mathrm{ZZ}}$. *Then*

$$\mathrm{E}[\hat{\theta}_L] = \theta$$

$$\mathrm{Var}[\hat{\theta}_L] = \frac{n-2}{n-q-2}(1 - R^2)\mathrm{Var}[\hat{\theta}_C]$$

$$\mathrm{E}[S_L^2] = \mathrm{Var}[\hat{\theta}_L]$$

and

$$\Pr\{|\hat{\theta}_L - \theta| \leq H_L\} = 1 - \alpha$$

where $R^2 = \boldsymbol{\sigma}_{Y\mathrm{C}}\boldsymbol{\Sigma}_{\mathrm{CC}}^{-1}\boldsymbol{\sigma}_{\mathrm{C}Y}/\sigma_Y^2$, *the square of the multiple correlation coefficient.*

Under the assumptions of Theorem 1, the CV point estimator is unbiased, the associated variance estimator is unbiased, and the confidence interval achieves the nominal coverage level. If $R^2 > q/(n-2)$, then $\mathrm{Var}[\hat{\theta}_L] < \mathrm{Var}[\hat{\theta}_C]$. Our goal is to examine the consequences of, and remedies for, violation of the assumption on which this result is based.

There is a close connection between least-squares regression and CVs that makes the reason we refer to $\hat{\theta}_L$ as the linear CV apparent. Let $\mathbf{X}$ be the $n \times (q+1)$ matrix whose ith row is $\mathbf{X}_i' = (1, (\mathbf{C}_i - \boldsymbol{\mu})')$. Then $\hat{\theta}_L$ is the first element of the $(q+1) \times 1$ vector

$$\hat{\gamma} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{G}\mathbf{X}'\mathbf{Y}$$

and $S_L^2 = S^2 G_{11}$, where G_{11} is the 11-element of $\mathbf{G}$ and S^2 is the residual variance defined above.

Stated differently, $\hat{\theta}_L$ is the estimator of the intercept term in the least-squares regression of Y_i on $\mathbf{C}_i - \boldsymbol{\mu}$. This makes sense because, under the assumption of multivariate normality, Y_i can be represented as

$$Y_i = \theta + (\mathbf{C}_i - \boldsymbol{\mu})'\boldsymbol{\beta} + \epsilon_i$$

where $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathrm{CC}}^{-1}\boldsymbol{\sigma}_{\mathrm{C}Y}$, and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. normal random variables with mean 0, variance $(1 - R^2)\sigma_Y^2$, and are independent of $\mathbf{C}$. This representation suggests that some of the results of Theorem 1 will hold under weaker assumptions analogous to those of least-squares regression when $\mathbf{X}$ is a fixed design matrix. Lavenberg, Moeller and Welch (1982) state the following results, which we prove in Nelson (1988); see also Cheng (1978).

Theorem 2. *Let* γ *be a* $(q+1) \times 1$*-vector of constants. If*: i) $\mathrm{E}[Y_i \mid \mathbf{C}_i] = \mathbf{X}_i'\boldsymbol{\gamma}$, *then* $\mathrm{E}[\hat{\theta}_L] = \theta$. *If, in addition,* ii) $\mathrm{Var}[Y_i \mid \mathbf{C}_i] = \sigma^2$ *independent of* $\mathbf{C}_i$, *then* $\mathrm{E}[S^2 G_{11}] = \mathrm{Var}[\hat{\theta}_L]$, $\gamma' = (\theta, \boldsymbol{\beta}')$ *and* $\sigma^2 = (1 - R^2)\sigma_Y^2$. *Finally, if it is also true that*: iii) *the conditional distribution of* Y_i *given* $\mathbf{C}_i$ *is univariate normal, then* $\Pr\{|\hat{\theta}_L - \theta| \leq H_L\} = 1 - \alpha$.

Theorem 2 shows that a linear regression implies that $\hat{\theta}_L$ is unbiased, the addition of constant conditional variance ensures that the variance estimator is also unbiased, and normality of the conditional distribution of Y leads to a valid confidence interval. Multivariate normality is sufficient, but not necessary, for these three conditions to hold. The following theorem states that, asymptotically, neither the assumptions of Theorem 1 nor 2 are required. A proof is given in Nelson (1988).

Theorem 3. *As* $n \to \infty$

$$\hat{\theta}_L \xrightarrow{P} \theta$$

$$nS_L^2 \xrightarrow{P} \sigma^2$$

and

$$\sqrt{n}(\hat{\theta}_L - \theta) \Rightarrow \mathrm{N}(0, \sigma^2)$$

where $\sigma^2 = (1 - R^2)\sigma_Y^2$, $\xrightarrow{P}$ *denotes convergence in probability,* $\Rightarrow$ *denotes weak convergence, and* $\mathrm{N}(0, \sigma^2)$ *denotes a normally distributed random variable with mean* 0 *and variance* σ^2.

Theorem 3 justifies using the linear CV point, variance and interval estimators when n is large. In contrast, $\hat{\theta}_C$ and S_C^2 are always unbiased, and the central limit theorem gives $\sqrt{n}(\hat{\theta}_C - \theta) \Rightarrow \mathrm{N}(0, \sigma_Y^2)$. Thus, if $R^2 > 0$, then $\hat{\theta}_L$ has an asymptotically smaller variance than $\hat{\theta}_C$.

How do departures from the assumptions of Theorem 2 affect $\hat{\theta}_L$ in small samples? We give two examples that relax conditions i and ii, respectively, replacing them with more general relationships. Although we cannot say whether these examples represent what would be encountered in practice, they do give an indication of the problems that might arise when the assumptions of Theorem 2 are not satisfied.

Suppose that

$$\mathrm{E}[Y_i \mid \mathbf{C}_i] = \eta + (\mathbf{C}_i - \boldsymbol{\mu})'\psi + \tfrac{1}{2}(\mathbf{C}_i - \boldsymbol{\mu})'\boldsymbol{\Delta}(\mathbf{C}_i - \boldsymbol{\mu}) \qquad (1)$$

where ψ is a $q \times 1$ vector and $\boldsymbol{\Delta}$ is a $q \times q$ symmetric matrix of constants. Then

$$\mathrm{E}[Y] = \eta + \tfrac{1}{2}\mathrm{trace}[\boldsymbol{\Delta}\boldsymbol{\Sigma}_{\mathrm{CC}}]$$

$$\mathrm{E}[\hat{\theta}_L] = \eta + \frac{1}{2}\left(\frac{n-2}{n}\right)\mathrm{trace}[\boldsymbol{\Delta}\boldsymbol{\Sigma}_{\mathrm{CC}}]$$

and

$$\mathrm{Bias}[\hat{\theta}_L] = -\frac{1}{n}\mathrm{trace}[\boldsymbol{\Delta}\boldsymbol{\Sigma}_{\mathrm{CC}}]$$

(Beale 1985). Thus, nonlinear regression leads to a biased CV point estimator. The bias is $O(n^{-1})$ when the true regression is quadratic, which could be severe in small samples.

Next, suppose that $q = 1$ and condition i of Theorem 2 holds, but $\mathrm{Var}[Y_i \mid C_i] = \upsilon C_i$, where υ is a constant; that is, the conditional variance is proportional to C_i. In Appendix B we show that

$$\mathrm{Var}[\hat{\theta}_L] = \mathrm{E}\left[\frac{\upsilon \bar{C}}{n}\right] + \mathrm{E}\left[\frac{(\bar{C} - \mu)^2}{S_{CC}} \frac{\upsilon \sum_{i=1}^{n} (C_i - \bar{C})^2 C_i}{(n-1)^2 S_{CC}}\right]. \qquad (2)$$

However,

$$\mathrm{E}[S_L^2] = \mathrm{E}\left[\left(\frac{n-1}{n} + \frac{(\bar{C}-\mu)^2}{S_{CC}}\right)\frac{\upsilon \bar{C}}{n-2}\right] - \frac{1}{n-2} \cdot \mathrm{E}\left[\left(\frac{n-1}{n} + \frac{(\bar{C}-\mu)^2}{S_{CC}}\right)\frac{\upsilon \sum_{i=1}^{n} (C_i - \bar{C})^2 C_i}{(n-1)^2 S_{CC}}\right]. \qquad (3)$$

Clearly S_L^2 is biased, but since $(n-1)/n + (\bar{C} - \mu)^2/S_{CC} \approx 1$ the first terms on the right-hand sides of (2) and (3) are nearly the same. Unfortunately, the second terms differ in sign for all values of n, and decrease in absolute value with n at different rates. Thus, nonconstant conditional variance leads to a biased CV variance estimator.

In the simulation study (Section 8) we include examples with nonlinear regression and nonconstant conditional variance.

The computation required to calculate the CV point, variance and interval estimators is clearly greater than the computation required for the corresponding crude estimators. The most significant additional calculation is $\mathbf{G} = (\mathbf{X}'\mathbf{X})^{-1}$. However, the difficulty of this calculation depends only on q, the number of controls, not on n, because $\mathbf{G}$ is $(q+1) \times (q+1)$. Typically q is small in the range of $1 \leq q \leq 5$.

3. BATCHING

For integers m and k such that $km = n$, define the jth *batch mean* to be

$$\bar{\mathbf{Z}}_j(k) = m^{-1} \sum_{i=(j-1)m+1}^{jm} \mathbf{Z}_i$$

for $j = 1, 2, \ldots, k$. The batch means are, of course, i.i.d. because so are the $\{\mathbf{Z}_i\}$. If the original process is normally distributed, then so are the batch means. However, when the original process is not normal, the batch means will tend to be more nearly normal as the batch size m increases (the number of batches k decreases). This property suggests forming the CV estimator from the k batch means $\bar{\mathbf{Z}}_j(k)$, $j = 1, 2, \ldots, k$.

Batching is a well known output analysis method for estimating the variance of the sample mean of a *dependent* output process (e.g., Schmeiser 1982). In that context, it is hoped that the batch means will more nearly be independent and normally distributed than the original process. Our concern here is only the nonnormality of $\{\mathbf{Z}_i\}$, which may result in $\hat{\theta}_L$ being biased.

The penalty for batching is a loss of degrees of freedom. Nelson (1989) quantified this penalty with respect to the performance of the point and interval estimator when the $\{\mathbf{Z}_i\}$ process is actually multivariate normal. Let $\hat{\theta}_B(k)$ be the linear CV point estimator based on k batch means of size $m = n/k$, so that $\hat{\theta}_L = \hat{\theta}_B(n)$.

Theorem 4. (Nelson 1989, Result 1) *Suppose that* $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ *are i.i.d.* $q+1$ *variate normal vectors. If* $0 < q < k < n$, *then*

$$\frac{\mathrm{Var}[\hat{\theta}_B(k)]}{\mathrm{Var}[\hat{\theta}_L]} = \frac{(k-2)(n-q-2)}{(n-2)(k-q-2)} > 1.$$

The ratio in Theorem 4 quantifies the loss in variance reduction due to batching when batching is not needed. An important conclusion is that the loss is negligible when $k > 60$ for $0 < q \leqslant 5$, no matter how large n is. Nelson (1989) also examined properties of the confidence interval and found that a similar statement holds: There is little penalty for batching provided that no fewer than 30 batches are used; up to 60 batches are worthwhile if as many as 5 controls are employed.

When the $\{\mathbf{Z}_i\}$ process is *not* multivariate normal, batching tends to reduce point-estimator and variance-estimator bias, and improves the coverage probability of the confidence interval, because the batch means are more nearly normal. Since the penalty for batching when it is not needed is negligible, provided the number of batches is not too small, it appears that the outputs should always be batched to between 30 and 60 batches when n is large. Within this range, the number of batches, k, should be selected so that it divides n evenly. The additional computation beyond the linear CV is slight: n vector additions and k vector divisions. The difficulty of calculating $\mathbf{G}$ is unaffected by batching.

4. KNOWN COVARIANCE CONTROLS

The coefficient $\hat{\boldsymbol{\beta}}$ that appears in the expression for $\hat{\theta}_L$ is an estimator of $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{CC}^{-1}\boldsymbol{\sigma}_{CY}$, the vector that minimizes $\mathrm{Var}[\hat{\theta}_C - (\overline{\mathbf{C}} - \boldsymbol{\mu})'\boldsymbol{\beta}]$ over all $q \times 1$ vectors $\boldsymbol{\beta}$, regardless of the assumptions that apply. Although $\boldsymbol{\sigma}_{CY}$ is rarely known in practice, $\boldsymbol{\Sigma}_{CC}$ is often known or can be calculated (e.g., Bauer, Venkatraman and Wilson 1987, and the M/M/1 example in Section 8). Bauer proposes a CV estimator that makes use of $\boldsymbol{\Sigma}_{CC}$

$$\hat{\theta}_K = \hat{\theta}_C - (\overline{\mathbf{C}} - \boldsymbol{\mu})'\ddot{\boldsymbol{\beta}}$$

where $\ddot{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{CC}^{-1}\mathbf{S}_{CY}$. The variance estimator is

$$S_K^2 = \frac{n-2}{n(n-1)} S_{Y \cdot C}^2 + \frac{q+1}{n(n-1)} S_Y^2$$

where $S_Y^2 = nS_C^2$ and

$$S_{Y \cdot C}^2 = \frac{n-1}{n-q-1} S_Y^2 \left(1 - \frac{\mathbf{S}_{YC}\mathbf{S}_{CC}^{-1}\mathbf{S}_{CY}}{S_Y^2}\right).$$

The associated confidence interval is $\hat{\theta}_K \pm H_K$, where $H_K = t_{\alpha/2}(n-q-1)S_K$. The following theorem summarizes the theoretical properties provided or implied by Bauer.

Theorem 5. (Bauer 1987) *Suppose that* $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ *are i.i.d.* $q+1$ *variate normal vectors with mean vector* $(\theta, \boldsymbol{\mu}')'$ *and variance* $\boldsymbol{\Sigma}_{ZZ}$. *Then*

$$\mathrm{E}[\hat{\theta}_K] = \theta$$

$$\mathrm{Var}[\hat{\theta}_K] = \frac{n+q-1}{n-1}\left(1 - \frac{n-2}{n+q-1} R^2\right)\mathrm{Var}[\hat{\theta}_C]$$

and

$$\mathrm{E}[S_K^2] = \mathrm{Var}[\hat{\theta}_K].$$

Notice that, under the assumptions of Theorem 5, $\mathrm{Var}[\hat{\theta}_L] < \mathrm{Var}[\hat{\theta}_K] < \mathrm{Var}[\hat{\theta}_C]$ if $R^2 > q/(n-2)$, the same requirement for a variance reduction as the linear CV. However, $\mathrm{Var}[\hat{\theta}_C] \leqslant \mathrm{Var}[\hat{\theta}_K] \leqslant \mathrm{Var}[\hat{\theta}_L]$ if $R^2 \leqslant q/(n-2)$.

The weaker assumption i of Theorem 2 is not sufficient to guarantee that $\hat{\theta}_K$ is unbiased. In fact, properties unique to the normal distribution are critical in the proof of unbiasedness. The confidence interval proposed by Bauer is approximate, even under the assumptions of Theorem 5. However, it was found to be robust in extensive experiments by Bauer (1987) and Bauer, Venkatraman and Wilson (1987). For that reason $\hat{\theta}_K$ was included in this study. The computation required is somewhat more than the linear CV because of the calculation of $\mathbf{S}_{Y \cdot C}$, which cannot be obtained directly from $\mathbf{G}$.

5. JACKKNIFING

Jackknifing is a well known resampling technique that is used to reduce point estimator bias, provide an estimate of standard error, and yield a robust confidence interval. For a review of jackknifing see Efron (1982) and Efron and Gong (1983); for a discussion of jackknifing in simulation see Lavenberg, Moeller and Welch (1982) and Bratley, Fox and Schrage (1987).

We begin with a traditional presentation of jackknifing as it applies to the control variate problem. Let $\hat{\theta}_L^{-i}$ be the linear CV computed without $\mathbf{Z}_i$; that is, the linear CV computed from $n-1$ replications, excluding the ith. The superscript $-i$ is used throughout to indicate the absence of the ith replication. Let $\tilde{\theta}_i = n\hat{\theta}_L - (n-1)\hat{\theta}_L^{-i}$, which is sometimes called the ith *pseudovalue*. The jackknife point and variance estimators are, respectively

$$\hat{\theta}_J = n^{-1} \sum_{i=1}^{n} \tilde{\theta}_i$$

and

$$S_J^2 = (n(n-1))^{-1} \sum_{i=1}^{n} (\tilde{\theta}_i - \hat{\theta}_J)^2$$

with associated confidence interval $\hat{\theta}_J \pm H_J$, where $H_J = t_{\alpha/2}(n-1)S_J$.

Before we establish properties of these estimators we present a computational result. Let $\boldsymbol{\delta}_i$ be a $(q+1) \times 1$ vector whose ith element is 1, and other elements are 0, and let $\mathbf{I}_{n\times n}$ be the $n \times n$ identity matrix. Let $\mathbf{H} = \mathbf{I}_{n\times n} - \mathbf{XGX}' = (H_{ij})$. Finally, let $\mathbf{V}$ be an $n \times 1$ vector whose ith element is

$$V_i = \frac{\boldsymbol{\delta}_1' \mathbf{GX}_i}{H_{ii}}.$$

In Appendix B we show that $\hat{\theta}_J = \hat{\theta}_L + ((n-1)/n)\mathbf{V}'\mathbf{HY} = \hat{\theta}_L + ((n-1)/n)\mathbf{V}'\hat{\boldsymbol{\epsilon}}$, where $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is the vector of residuals from the least-squares regression that forms $\hat{\theta}_L$. In terms of computation, the implication is that only one matrix inversion, $\mathbf{G}$, is required to compute $\hat{\theta}_J$.

The following theorem establishes the asymptotic normality of $\hat{\theta}_J$, which justifies, to some extent, using the t distribution to form confidence intervals. In the simulation literature the results cited to establish normality apply when $\mathbf{X}$ is a fixed design matrix (e.g., Miller 1974 and Hinkley 1977). While it is possible to prove the asymptotic normality of $\hat{\theta}_J$ using the results of Arvensen (1969) for U-statistics, the computational formula above permits a direct proof.

Theorem 6. *As* $n \to \infty$, $\sqrt{n}(\hat{\theta}_J - \theta) \Rightarrow \mathrm{N}(0, \sigma^2)$, *where* $\sigma^2 = (1 - R^2)\sigma_Y^2$.

The next theorem establishes some properties of $\hat{\theta}_J$ in small samples.

Theorem 7. *Suppose that condition* i *of Theorem* 2 *or Equation* (1) *holds. Then* $\mathrm{E}[\hat{\theta}_J] = \theta$. *If conditions* i *and* ii *hold, then*

$$\mathrm{Var}[\hat{\theta}_J] = \mathrm{Var}[\hat{\theta}_L] + \left(\frac{n-1}{n}\right)^2 \sigma^2 \mathrm{E}[\mathbf{V}'\mathbf{HV}]$$
$$\geqslant \mathrm{Var}[\hat{\theta}_L].$$

Theorem 7 shows that $\hat{\theta}_J$ is unbiased when the regression is linear or quadratic. In general, jackknifing eliminates $O(n^{-1})$ bias, which is precisely the quadratic case. Theorem 7 also shows that jackknifing tends to inflate point-estimator variance. We examine this inflation in more detail in Theorem 8, which quantifies the penalty for jackknifing when it is least needed.

Theorem 8. *Suppose that conditions* i *and* ii *of Theorem* 2 *hold. Then* $\mathrm{Var}[\hat{\theta}_J] - \mathrm{Var}[\hat{\theta}_L] = O(n^{-3})$.

Thus, as n increases, the variance inflation becomes small relative to the variance of $\hat{\theta}_J$.

Regarding the variance-estimator S_J^2, results due to Efron and Stein (1981) suggest that S_J^2 will tend to overestimate $\mathrm{Var}[\hat{\theta}_J]$.

6. SPLITTING

The linear CV estimator $\hat{\theta}_L = \hat{\theta}_C - (\overline{\mathbf{C}} - \boldsymbol{\mu})'\hat{\boldsymbol{\beta}}$ would be an unbiased estimator of θ—even in the absence of condition i of Theorem 2—if $\overline{\mathbf{C}}$ and $\hat{\boldsymbol{\beta}}$ were independent, which they are not in general. However, they are independent if $\hat{\boldsymbol{\beta}}$ is computed from a preliminary or pilot sample, rather than from $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$. Ripley (1987) analyzes this approach: If $\hat{\boldsymbol{\beta}}^*$ is the estimator of $\boldsymbol{\beta}$ based on an independent sample $\mathbf{Z}_1^*, \ldots, \mathbf{Z}_{n^*}^*$, then under the assumptions of Theorem 1

$$\mathrm{Var}[\hat{\theta}_C - (\overline{\mathbf{C}} - \boldsymbol{\mu})'\hat{\boldsymbol{\beta}}^*]$$
$$= \left(\frac{n^* - 2}{n^* - q - 2}\right)(1 - R^2)\mathrm{Var}[\hat{\theta}_C].$$

Even if a pilot experiment is not feasible, Ripley's analysis suggests that, when n is quite large, a reasonable strategy might be to dedicate a portion of the sample to estimate $\boldsymbol{\beta}$.

A related approach, which was first suggested by Tocher (1963), is to split the sample $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ into two or more groups, compute an estimate of $\boldsymbol{\beta}$ from each group, then exchange the estimates among the groups. The primary difficulty with this approach is to compute a variance estimate. In this section, we consider an extreme form of splitting—splitting into n groups—that permits a variance estimate and approximate confidence interval. Our point estimator is similar to Beale (1985).

The splitting estimator is

$$\hat{\theta}_S = n^{-1} \sum_{i=1}^{n} (Y_i - (\mathbf{C}_i - \boldsymbol{\mu})'\hat{\boldsymbol{\beta}}^{-i}) = n^{-1} \sum_{i=1}^{n} \overline{\theta}_i$$

with variance estimator

$$S_S^2 = (n(n-1))^{-1} \sum_{i=1}^{n} (\overline{\theta}_i - \hat{\theta}_S)^2$$

and associated confidence interval $\hat{\theta}_S \pm H_S$, where

$H_S = t_{\alpha/2}(n-1)S_S$ and $-i$ indicates calculations without $\mathbf{Z}_i$. Clearly $\hat{\theta}_S$ is unbiased, and S_S^2 is the appropriate variance estimator under the (incorrect) assumption that the $\{\overline{\theta}_i\}$ are independent. We establish additional properties below.

Let $\mathbf{W}$ be an $n \times 1$ vector whose ith element is

$$W_i = \frac{(\mathbf{X}_i - \boldsymbol{\delta}_1)'\mathbf{G}\mathbf{X}_i}{H_{ii}}.$$

In Appendix B we obtain a computational result that is analogous to jackknifing: $\hat{\theta}_S = \hat{\theta}_L + n^{-1}\mathbf{W}'\mathbf{HY} = \hat{\theta}_L + n^{-1}\mathbf{W}'\hat{\boldsymbol{\epsilon}}$. Thus, the computation required to calculate $\hat{\theta}_S$ is about the same as $\hat{\theta}_J$. This formula can be used to establish properties analogous to those established for $\hat{\theta}_J$:

Theorem 9. *As* $n \to \infty$, $\sqrt{n}(\hat{\theta}_S - \theta) \Rightarrow \mathrm{N}(0, \sigma^2)$, *where* $\sigma^2 = (1 - R^2)\sigma_Y^2$.

Theorem 10. *Suppose that conditions* i *and* ii *of Theorem* 2 *hold, then*

$$\mathrm{Var}[\hat{\theta}_S] = \mathrm{Var}[\hat{\theta}_L] + \frac{\sigma^2}{n^2}\,\mathrm{E}[\mathbf{W}'\mathbf{HW}] \geqslant \mathrm{Var}[\hat{\theta}_L].$$

Like the jackknife, splitting inflates the variance. The following theorem shows that, as n increases, the inflation from splitting is of the same order as jackknifing. However, the simulation results reported later indicate that $\hat{\theta}_S$ is superior to $\hat{\theta}_J$ for small n.

Theorem 11. *Suppose that conditions* i *and* ii *of Theorem* 2 *hold. Then* $\mathrm{Var}[\hat{\theta}_S] - \mathrm{Var}[\hat{\theta}_L] = O(n^{-3})$.

The next result justifies using S_S^2 to estimate $\mathrm{Var}[\hat{\theta}_S]$.

Theorem 12. *As* $n \to \infty$, $nS_S^2 \xrightarrow{P} \sigma^2 = (1 - R^2)\sigma_Y^2$. *If conditions* i *and* ii *of Theorem* 2 *hold, then*

$$\mathrm{E}[S_S^2] = \left(\frac{n+q-1}{n-1}\right)\frac{\sigma^2}{n} + \frac{\sigma^2}{n^2}\,\mathrm{E}[\mathbf{W}'\mathbf{HW}] - \frac{2\sigma^2}{n(n-1)}\sum_{i \neq j}\mathrm{E}[W_i H_{ij} W_j].$$

The first part of Theorem 12 shows that S_S^2 is a reasonable estimator in large samples. The second part is a small-sample result for a special case. If we also assume that the $[\mathbf{Z}_i\}$ are normally distributed, then Theorems 1 and 10 imply that

$$\mathrm{Var}[\hat{\theta}_S] = \left(\frac{n-2}{n-q-2}\right)\frac{\sigma^2}{n} + \frac{\sigma^2}{n^2}\,\mathrm{E}[\mathbf{W}'\mathbf{HW}].$$

We can show that $(n+q-1)/(n-1) < (n-2)/(n-q-2)$ when $q > 0$, which means that S_S^2 underestimates the leading term in $\mathrm{Var}[\hat{\theta}_S]$ in small samples. The simulation results bear out this tendency.

7. BOOTSTRAPPING

For completeness, we describe how bootstrapping (Efron 1982) can be used to form a CV estimator. However, we also discuss why we did not include the bootstrap in the full simulation study.

Let $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ be a realization of $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$, and let $\hat{F}$ be the probability distribution that puts mass $1/n$ on each of $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$. Consider drawing an i.i.d. sample of size n with replacement from $\hat{F}$. If we let a subscript $\hat{F}$ denote calculations with respect to such sampling, then the bootstrap point estimator is $E_{\hat{F}}[\hat{\theta}_L]$, with associated variance estimator $\mathrm{Var}_{\hat{F}}[\hat{\theta}_L]$. Here $\hat{\theta}_L$ is the linear CV computed from a sample of size n from $\hat{F}$.

Directly calculating these estimates is computationally expensive when n is large because there are $\binom{2n-1}{n}$ distinct samples of size n from $\hat{F}$, and the linear CV must be computed for each sample. The standard approximation is to draw $b > n$ random samples of size n from $\hat{F}$ and let estimates replace the calculations.

Let $\hat{\theta}_L^{\ell}$ be the linear CV computed from the ℓth bootstrap sample of size n, for $\ell = 1, 2, \ldots, b$. Then the bootstrap point and variance estimators are, respectively

$$\hat{\theta}_E = b^{-1}\sum_{i=1}^{b}\hat{\theta}_L^{\ell}$$

and

$$S_E^2 = (b-1)^{-1}\sum_{\ell=1}^{n}(\hat{\theta}_L^{\ell} - \hat{\theta}_E)^2$$

(E stands for Efron). As $b \to \infty$ these sample versions converge to the bootstrap estimates defined above.

Computational expense is one obvious drawback of bootstrapping for CVs, even when exact calculations are replaced by sampling. Bootstrapping requires the equivalent of drawing b samples from a multinomial distribution with n equiprobable cells, and then computing b linear CV estimators $\hat{\theta}_L^{\ell}$. Efron recommends *b in the range of 50 to 200.*

An additional problem is to construct a confidence interval. The theory of bootstrap interval estimation is still developing, and there is currently no standard procedure. A crude confidence interval is $\hat{\theta}_E \pm z_{\alpha/2}S_E$, where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. A different interval can be formed by estimating the $\alpha/2$ and $1 - \alpha/2$ quantiles of the random variable $\hat{\theta}_L$ from the b i.i.d. bootstrap estimates $\hat{\theta}_L^1, \ldots, \hat{\theta}_L^b$; e.g., the interval $(\hat{\theta}_L^{(l)}, \hat{\theta}_L^{(u)})$ where $l = \lfloor b\alpha/2 \rfloor + 1$, $u = \lfloor b(1 - \alpha/2) \rfloor + 1$, and $\hat{\theta}_L^{(\ell)}$ is the ℓth largest value of $\hat{\theta}_L^1, \ldots, \hat{\theta}_L^b$. Refinements of this interval have been proposed (e.g., Efron and Tibshirani 1986). However, it is apparent that an even larger number of bootstrap samples is needed for interval estimation than for point and variance estimation.

Since computational expense makes it impractical, and since there is not yet a satisfactory interval estimator, we did not include $\hat{\theta}_E$ in the full simulation study reported below. We did perform some experiments and the results are summarized.

8. SIMULATION STUDY

In this section, we summarize the results of an extensive simulation study of the estimators described earlier. Rather than use only systems simulation examples, we also chose several *models* of simulation output processes. These models represent factors that affect CV performance, including: the joint distribution of $\mathbf{Z}$; the marginal distribution of Y; the regression of Y on $\mathbf{C}$, $\mathrm{E}[Y \mid \mathbf{C}]$; the conditional variance of Y given $\mathbf{C}$, $\mathrm{Var}[Y \mid \mathbf{C}]$; the sample size, n, and the number of controls, q; and the squared multiple correlation coefficient, R^2. In all cases, models were selected for which θ is known so that confidence interval coverage could be evaluated. The models, the experiment design, and the results are described below. All experiments were performed on a Pyramid 90x supermini computer, and all analysis, including graphics, was done using the S-system (Becker and Chambers 1984). Random variates were generated using either the IMSL library, version 10, with the default constants for random number generation, or the S-system.

8.1. Models

To investigate the penalty for applying CV remedies when they are not needed, the multivariate normal distribution was used as a model. Multivariate normal vectors were generated using IMSL routines chfac and rnmvn to compute the Cholesky decomposition of $\mathbf{\Sigma}_{ZZ}$ and to generate the multivariate normal variates, respectively.

The multivariate Pearson Type VII distribution (Johnson 1987) is an elliptically contoured distribution with a linear regression, both properties shared by the normal distribution. However, unlike the normal distribution, $\mathrm{Var}[Y \mid \mathbf{C}]$ is not constant. The variance-covariance matrix can be specified, and the shape of the marginal distributions can be controlled through a single parameter, ω. A FORTRAN program was written to implement the algorithm of Johnson (p. 118). IMSL routines chfac, rnmvn, and rnchi were used for variate generation.

Plackett's bivariate distribution (Johnson 1987) also has a linear regression, but the marginals are uniform on the interval (0, 1). Other marginals can be obtained through the method of inversion; we used exponential marginals with mean 1. Dependence is controlled through a single parameter, Ψ. A FORTRAN program was written to implement the algorithm of Johnson (p. 193); IMSL routine rnun was used to generate uniform (0, 1) random variates.

To obtain more direct control over the factors of interest, we extended the "functional approach" of Kottas and Lau (1978) for bivariate random variables to multivariate random variables; we call this the extended KL distribution. Let $\mathbf{C}_i$ be a $q \times 1$ random vector having distribution $\mathrm{N}(\mathbf{0}_{q\times 1}, \mathbf{\Sigma}_{CC})$, where $\mathbf{0}_{q\times 1}$ is a $q \times 1$ vector of zeros, and let ϵ_i be a random variable with mean 0 and variance 1 that is independent of $\mathbf{C}_i$. Then for scalar η, $q \times 1$ vectors $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$, and $q \times q$ symmetric matrix $\mathbf{\Delta}$, all constants, let

$$Y_i = \eta + \mathbf{C}_i'\boldsymbol{\psi} + \mathbf{C}_i'\mathbf{\Delta}\mathbf{C}_i + \epsilon_i\mathbf{C}_i'\boldsymbol{\phi}.$$

If we further restrict $\mathbf{\Delta}$ to be diagonal, and let the operator *sum* add all the elements in a matrix, then the following properties can be derived by direct calculation

$$\mathrm{E}[Y_i] = \eta + \mathrm{trace}[\mathbf{\Delta}\mathbf{\Sigma}_{CC}]$$

$$\mathrm{E}[Y_i \mid \mathbf{C}_i = \mathbf{c}_i] = \eta + \mathbf{c}_i'\boldsymbol{\psi} + \mathbf{c}_i'\mathbf{\Delta}\mathbf{c}_i$$

$$\mathrm{Var}[Y_i] = \boldsymbol{\phi}'\mathbf{\Sigma}_{CC}\boldsymbol{\phi} + \boldsymbol{\psi}'\mathbf{\Sigma}_{CC}\boldsymbol{\psi} + 2\,\mathrm{sum}[\mathbf{\Delta}\mathbf{\Sigma}_{CC}\mathbf{\Delta}]$$

$$\mathrm{Var}[Y_i \mid \mathbf{C}_i = \mathbf{c}_i] = (\mathbf{c}_i'\boldsymbol{\phi})^2$$

and

$$\mathrm{Cov}[\mathbf{C}_i, Y_i] = \mathbf{\Sigma}_{CC}\boldsymbol{\psi}.$$

The regression of Y on $\mathbf{C}$ and the conditional variance are quadratic. The conditional distribution of Y is determined by the marginal distribution of ϵ; e.g., if ϵ is normally distributed then the conditional distribution of Y is normal. When designing experiments $\mathbf{\Sigma}_{CC}$ can be specified directly and R^2 controlled indirectly through the parameters $\boldsymbol{\psi}$, $\mathbf{\Delta}$ and $\boldsymbol{\phi}$. For example, if $\mathbf{\Sigma}_{CC} = \mathbf{I}_{q\times q}$, then

$$R^2 = \frac{\boldsymbol{\psi}'\boldsymbol{\psi}}{\boldsymbol{\psi}'\boldsymbol{\psi} + \boldsymbol{\phi}'\boldsymbol{\phi} + 2\,\mathrm{trace}[\mathbf{\Delta}^2]}.$$

The controls, $\mathbf{C}$, were generated using the S-function rnnorm, and ϵ was generated using either S-function rnorm or rgamma.

The distribution-sampling models above were selected to test the linear CV and remedies over the factors that affect their performance, factors that are usually not controllable in systems simulation examples. However, three systems simulation examples were also selected. The selection was based on common use, rather than because they are representative of real problems. Although these examples are less controllable than the distribution sampling models, they may include features not captured in those models.

The system simulation examples are a classical machine-repair system, an (s, S) inventory system, and an M/M/1 queue. Only results for the queueing system are reported here. The arrival rate to the queue is 0.9 customers/unit time and the service rate is 1 customer/unit time. Let θ be the expected delay in the queue for the 10th arriving customer. The four controls are: the sum of the interarrival times of the first 10 customers, the sum of the service times of the first 9 customers, the interarrival time of the 10th customer, and the service time of the 9th customer. The Var[**C**] can be calculated directly for these controls, so $\hat{\theta}_K$ can be employed. IMSL routine rnexp was used to generate exponential random variates.

8.2. Experiment Design

Large data sets of i.i.d. vectors were generated from the models just described. An *experiment*, as defined here, used the first mn vectors of a data set to compute m realizations of the six estimators $\hat{\theta}_C$, $\hat{\theta}_L$, $\hat{\theta}_B(k)$, $\hat{\theta}_K$, $\hat{\theta}_J$ and $\hat{\theta}_S$, and their associated variance and interval estimators, each based on a sample of size n; in some cases $\hat{\theta}_E$ was also computed. We call the m subsets of n vectors *macroreplications*, and the n vectors within a subset *microreplications*.

In initial experiments, $m = 100$ macroreplications were used for reasons described below; some follow-up experiments set $m = 400$. The number of microreplications was $n = 10$, 30 or 60. When $n = 10$ there are at least 4 degrees of freedom for each estimator. Sample sizes larger than $n = 60$ were not used because the results in Section 3 imply that larger samples should first be batched to form 30 to 60 batch means. For $\hat{\theta}_B(k)$ we set $k = 10$, 10 and 30 when $n = 10$, 30 and 60, respectively. The number of controls ranged from $q = 1, 2, \ldots, 5$, which was chosen by a subjective judgment that there are rarely more than five effective controls in a simulation.

The number of macroreplications, m, was chosen to allow confidence interval and variance comparisons. In all experiments, the nominal confidence level was $1 - \alpha = 0.95$. If the interval estimator's true coverage probability is close to the nominal level, then the standard error of the *estimated* coverage, based on m macroreplications, is $\sqrt{(0.95)(0.05)/m}$, which is approximately 0.023 when $m = 100$ and approximately 0.011 when $m = 400$. This was considered adequate resolution, and it provides a guide for deciding which observed differences are significant.

Because CVs is a variance reduction technique, point-estimator variance comparisons are especially important. Rather than compare the values of the variance estimators (e.g., S_L^2 versus S_J^2), which may be biased, we used an unbiased estimator of variance. To be specific, if $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are the m values of an estimator $\hat{\theta}$, then we estimate $\text{Var}[\hat{\theta}]$ *for the purpose of comparison* by $S_{\hat{\theta}}^2 = (m-1)^{-1} \sum_{i=1}^{m} (\hat{\theta}_i - \bar{\hat{\theta}})^2$, where $\bar{\hat{\theta}}^2 = m^{-1} \sum_{i=1}^{m} \hat{\theta}_i$.

If the $\{Y_i\}$ are normally distributed, then

$$\text{Var}[S_{\hat{\theta}_C}^2] = 2\,\frac{\text{Var}[\hat{\theta}_C]^2}{m-1} = \frac{2\sigma_Y^4}{n^2(m-1)}$$

implying that the coefficient of variation of $S_{\hat{\theta}_C}^2$ is $\sqrt{2/(m-1)}$, or approximately 0.14 when $m = 100$ and 0.07 when $m = 400$. We expect the variance of the CV estimators to be smaller than the variance of $\hat{\theta}_C$, so $\text{Var}[S_{\hat{\theta}}^2]$ for the CV estimators should be smaller as well. Thus, designing experiments based on the crude estimator is conservative. The coefficient of variation was used as a guide to determine the number of significant digits to report in Table II below, and this resolution was considered adequate to evaluate differences, particularly if we also consider the favorable effect of comparing estimators computed from the same data set.

Recall that

$$\mathbf{\Sigma}_{\mathbf{ZZ}} = \begin{pmatrix} \sigma_Y^2 & \boldsymbol{\sigma}_{YC} \\ \boldsymbol{\sigma}_{CY} & \mathbf{\Sigma}_{\mathbf{CC}} \end{pmatrix}.$$

As far as possible we controlled experimental conditions across models. The following covariance matrices were used for several models

$$\mathbf{\Sigma}_{\mathbf{ZZ}}^a = \begin{pmatrix} 1 & 0.7 & 0.5 & 0.3 & 0.2 & 0.1 \\ 0.7 & 1 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 1 & 0 & 0 & 0 \\ 0.3 & 0 & 0 & 1 & 0 & 0 \\ 0.2 & 0 & 0 & 0 & 1 & 0 \\ 0.1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{\Sigma}_{\mathbf{ZZ}}^b = \begin{pmatrix} 1 & 0.7 & 0.5 & 0.3 & 0.2 & 0.1 \\ 0.7 & 1 & 0.2 & 0.1 & 0 & 0 \\ 0.5 & 0.2 & 1 & 0.08 & 0 & 0 \\ 0.3 & 0.1 & 0.08 & 1 & 0 & 0 \\ 0.2 & 0 & 0 & 0 & 1 & 0 \\ 0.1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{\Sigma}^{c}_{ZZ} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

$$\mathbf{\Sigma}^{d}_{ZZ} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}.$$

In $\mathbf{\Sigma}^{a}_{ZZ}$ the controls are mutually independent and $R^2 = 0.88$. In $\mathbf{\Sigma}^{b}_{ZZ}$ some of the controls are dependent and $R^2 = 0.72$. The covariance vector, $\boldsymbol{\sigma}_{YC} = (0.7, 0.5, 0.3, 0.2, 0.1)$, was selected to examine the effect of adding controls from the most to least effective.

Table I summarizes the experimental cases for the distribution-sampling models. A boldface constant with a subscript denotes a matrix with all elements that are constant. For the Pearson examples the value of ω is the smallest value that permits $\mathbf{\Sigma}_{ZZ}$ to be specified; we want ω small because as ω increases the Pearson Type VII approaches the multivariate normal. In the first two extended KL examples only the covariance matrix of the controls could be matched with the normal and Pearson examples; $\mathbf{\Sigma}^{a}_{CC}$ implies $R^2 = 0.78$, and $\mathbf{\Sigma}^{b}_{CC}$ implies $R^2 = 0.81$. In the last four bivariate KL examples $R^2 = 0.5$. In all the KL examples ϵ was modeled as a N(0, 1) and as an $\mathscr{E}(1) - 1$ random variable, where $\mathscr{E}(1)$ is a random variable with an exponential distribution and mean 1. For the Plackett example $R^2 \approx 0.84$ for uniform marginals and 0.76 for exponential marginals.

8.3. Results

We do not attempt to present detailed results from such a large simulation study. Instead, we summarize general conclusions based on studying all the results, and present details of some typical examples.

8.3.1. Point Estimators

All the point estimators performed similarly in terms of variance, bias and mean squared error when the sample size reached $n = 60$. There were clear differences in many data sets when $n = 10$, which often persisted when $n = 30$.

The estimated MSE$[\hat{\theta}_L]$ was less than MSE$[\hat{\theta}_C]$ in all the data sets at all sample sizes. However, Bias$[\hat{\theta}_L]$ was significant for the extended KL examples with quadratic regression and the M/M/1 queue example, particularly at $n = 10$. When bias was present for $\hat{\theta}_L$, it was also present to the same degree in $\hat{\theta}_K$; however, in cases with nonnormal data, $\hat{\theta}_K$ sometimes showed bias when $\hat{\theta}_L$ was unbiased. In all the examples with nonlinear regression $\hat{\theta}_J$ showed no bias; recall that $\hat{\theta}_J$ is unbiased when the regression is quadratic. Of course, $\hat{\theta}_S$ is always unbiased.

Variance comparisons were surprisingly consistent across examples, with Var$[\hat{\theta}_L]$ always less than Var$[\hat{\theta}_C]$. The Var$[\hat{\theta}_B(k)]$ was sometimes significantly larger than the Var$[\hat{\theta}_L]$ when $n = 30$ and $k = 10$, but negligibly different when $n = 60$ and $k = 30$. The Var$[\hat{\theta}_K]$ was always greater than or equal to Var$[\hat{\theta}_L]$, and seemed to be much larger in the examples with nonconstant conditional variance. In some cases (see the example below), the variance of $\hat{\theta}_K$ was significantly larger than the variance of the crude estimator, and $\hat{\theta}_K$ was often beaten by $\hat{\theta}_J$ and $\hat{\theta}_S$.

At all sample sizes it appeared that Var$[\hat{\theta}_S] \leqslant$ Var$[\hat{\theta}_J]$, and at $n = 10$ the inequality was nearly always strict. The Var$[\hat{\theta}_S]$ typically was less than Var$[\hat{\theta}_C]$, and had the same variance as $\hat{\theta}_L$ for $n = 30$ or 60.

The effect of changing q, which we examined by adding controls one-at-a-time from most to least effective, was the same for all CV estimators.

Figures 1, 2 and 3 show box plots of $m = 100$ realizations of each estimator for samples of size $n = 10$ and 60. The box in a box plot contains the middle half of the data (0.25 to 0.75 sample quantiles), the center line is the median, and the whiskers extend to ±1.5 times the interquartile range beyond the median, or to the most extreme value, which ever is least; an asterisk indicates a value beyond ±1.5 times the

Table I
Experiment Cases for Distribution-Sampling Models

Distribution	Covariance Matrices	Additional Parameters
Normal	$\mathbf{\Sigma}^{x}_{ZZ}$, $x = a, b, c, d$	
Pearson VII	$\mathbf{\Sigma}^{x}_{ZZ}$, $x = a, b$	$\omega = 5$
Pearson VII	$\mathbf{\Sigma}^{x}_{ZZ}$, $x = c, d$	$\omega = 3$
Extended KL	$\mathbf{\Sigma}^{x}_{CC}$, $x = a, b$	$\eta = -5,\quad \boldsymbol{\psi}' = (1, 2, 3, 4, 5),\quad \boldsymbol{\Delta} = \mathbf{I}_{5\times5},\quad \boldsymbol{\phi} = \mathbf{1}_{5\times1}$
Extended KL	$\mathbf{\Sigma}^{x}_{CC}$, $x = a, b$	$\eta = 0,\quad \boldsymbol{\psi}' = (1, 2, 3, 4, 5),\quad \boldsymbol{\Delta} = \mathbf{0}_{5\times5}, \boldsymbol{\phi} = \sqrt{\mathbf{3}}_{5\times1}$
Extended KL		$\eta = 0,\quad \psi = \sqrt{1/2},\quad \Delta = 0,\quad \phi = \sqrt{1/2}$
Extended KL		$\eta = -\sqrt{1/8},\quad \psi = \sqrt{1/2},\quad \Delta = \sqrt{1/8},\quad \phi = 1/2$
Extended KL		$\eta = -1/2,\quad \psi = \sqrt{1/2},\quad \Delta = 1/2,\quad \phi = 0$
Extended KL		$\eta = -\sqrt{3/16},\quad \psi = \sqrt{1/2},\quad \Delta = \sqrt{3/16},\quad \phi = \sqrt{1/8}$
Plackett		$\boldsymbol{\Psi} = 30$

interquartile range. A line has been drawn across the entire plot at height θ.

Figure 1 shows the results for multivariate normal data with covariance matrix $\mathbf{\Sigma}^a_{ZZ}$. This model is one for which none of the remedies is needed. The variance inflation from jackknifing, relative to $\hat{\theta}_L$, is apparent at the smaller sample size. Although $\hat{\theta}_L$ is superior at both sample sizes, $\hat{\theta}_S$ is close.

Figure 2 shows an analogous plot for Pearson Type VII data with the same covariance structure as the data in Figure 1. This model has linear regression, but nonconstant conditional variance. The estimator $\hat{\theta}_K$ performs poorly in this example, as it did in several examples with nonconstant conditional variance.

The extreme value of $\hat{\theta}_K$ in Figure 2a is worthy of discussion. Because $\mathbf{\Sigma}^a_{CC} = \mathbf{I}_{5\times5}$, $\ddot{\boldsymbol{\beta}} = \mathbf{S}_{CY}$. Examination of the data set that yielded the extreme value revealed one vector whose elements were exceptionally large in absolute value. This resulted in the elements of $\mathbf{S}_{CY}$ being different from $\boldsymbol{\sigma}_{CY}$ by an order of magnitude. The same effect was present in $\mathbf{S}_{CC}$, but was a compensating effect for $\hat{\boldsymbol{\beta}}$ since $\hat{\boldsymbol{\beta}} = \mathbf{S}_{CC}^{-1}\mathbf{S}_{CY}$. Box plots for the same model with $m = 400$ macroreplications showed several more extreme points for $\hat{\theta}_K$. This result contradicts the conventional wisdom that variance is always reduced by incorporating all available information.

Figure 3 shows analogous plots for the first extended KL example in Table I with $\mathbf{\Sigma}^a_{CC}$ the covariance matrix of the controls and ϵ normally distributed. In addition to nonconstant conditional variance, the regression is quadratic in this model. The resulting bias is apparent for $\hat{\theta}_L$, $\hat{\theta}_B(k)$ and $\hat{\theta}_K$.

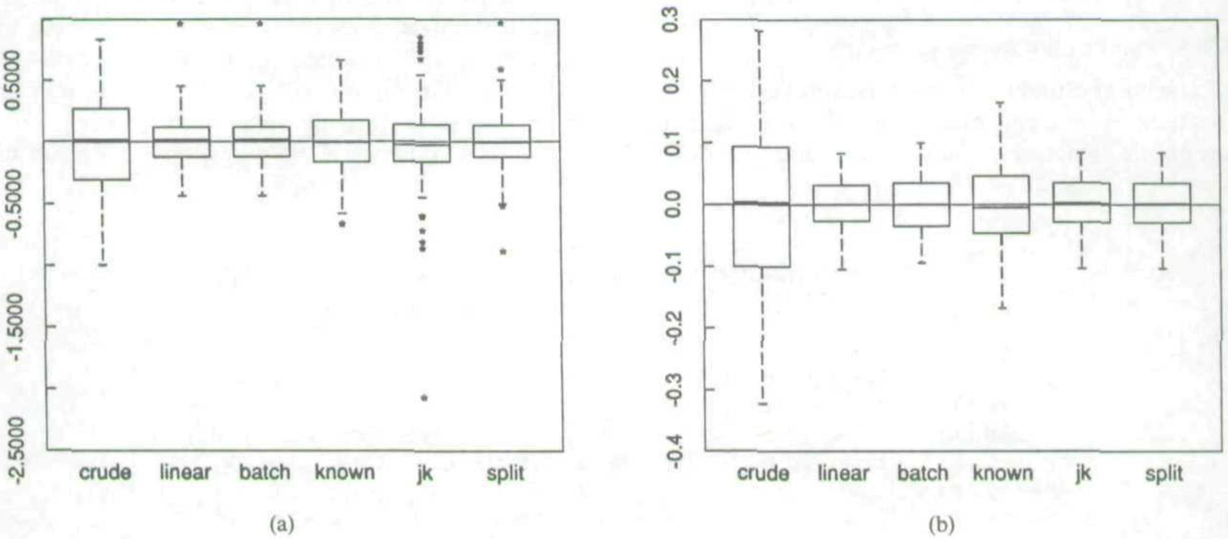
Table II gives complete numerical results for these three examples and the M/M/1 queue, including estimated MSE, variance, bias, expected halfwidth, expected value of the variance estimator, and coverage.

8.3.2. Variance Estimators

In each experiment, we compared the m variance estimators to $S^2_{\hat{\theta}}$ to assess how well each one estimated $\mathrm{Var}[\hat{\theta}]$, where $\hat{\theta}$ is generic for any of the six estimators. The estimator S^2_L, and thus $S^2_{B(k)}$, performed well throughout. When n was small, S^2_J tended to overestimate $\mathrm{Var}[\hat{\theta}_J]$ substantially, while S^2_S tended to underestimate $\mathrm{Var}[\hat{\theta}_S]$ slightly. In the data sets where $\mathrm{Var}[\hat{\theta}_K]$ was large relative to the other estimators, S^2_K underestimated the true variance. Some of these traits are apparent in the confidence interval comparisons below.

8.3.3. Confidence Intervals

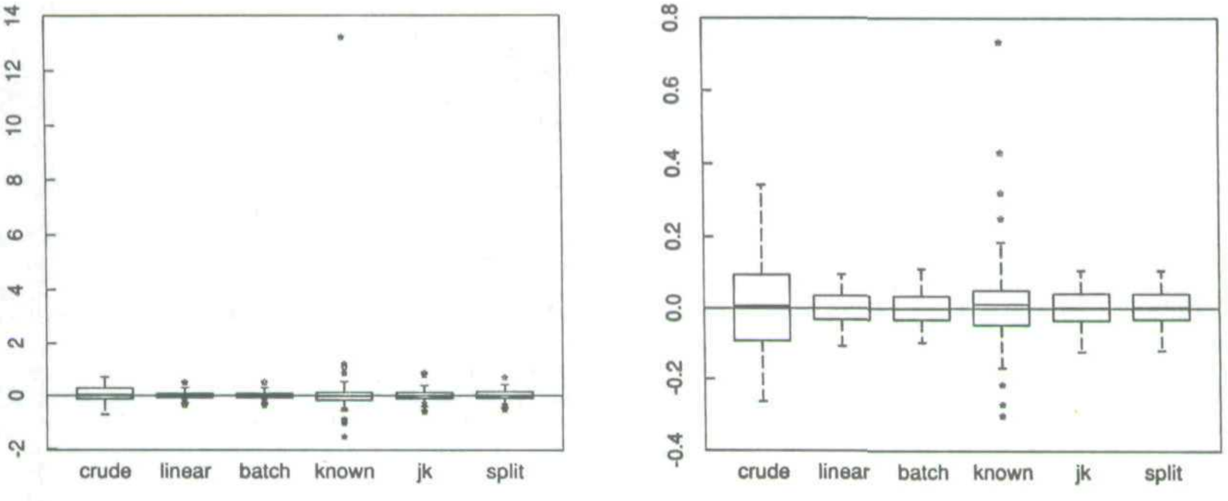
Point-estimator bias from nonlinear regression was the most significant factor that affected the coverage probability for $\hat{\theta}_L$, but even that was not a major problem when $n = 60$. Batching improved coverage somewhat. The confidence intervals associated with $\hat{\theta}_K$ and $\hat{\theta}_J$ often showed robust coverage, but their halfwidths were consistently larger than the other interval estimators. The coverage probability for



(a)

(b)

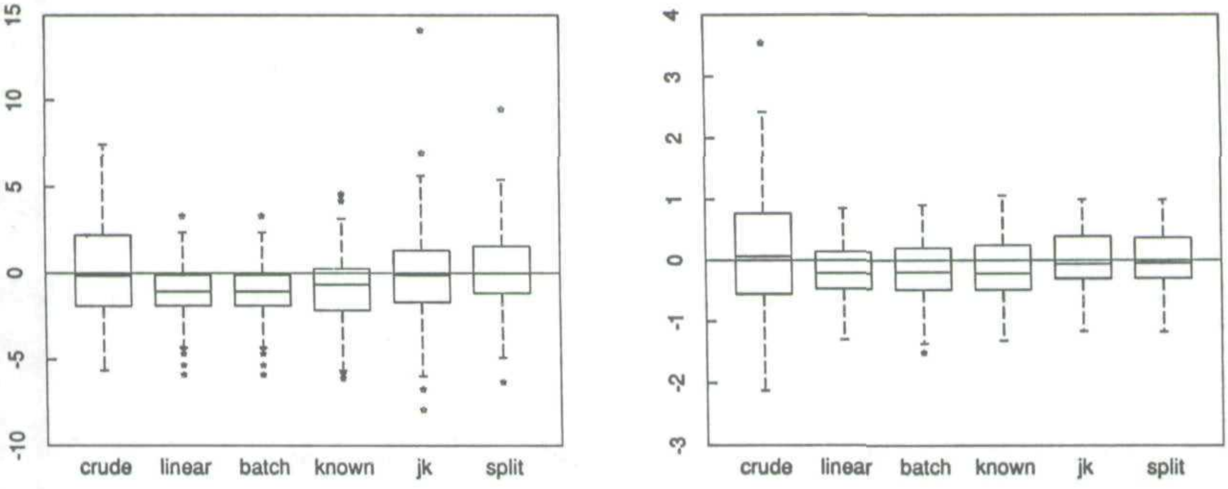
Figure 1. Box plots of point estimators for multivariate normal data with $\mathbf{\Sigma}^a_{ZZ}$ and a) $n = 10$, b) $n = 60$.



(a)

(b)

Figure 2. Box plots of point estimators for Pearson type VII data with $\boldsymbol{\Sigma}^{a}_{ZZ}$ and a) $n = 10$, b) $n = 60$.



(a)

(b)

Figure 3. Box plots of point estimators for extended KL data with $\boldsymbol{\Sigma}^{a}_{CC}$ and a) $n = 10$, b) $n = 60$.

the interval associated with $\hat{\theta}_S$ was often lower than the nominal level when $n = 10$, typically about 0.88. This can be explained partly by underestimation of the variance, as mentioned above. The effect of nonnormal marginal distributions was only apparent when n was small.

Figures 4, 5 and 6 show midpoint by halfwidth plots for pairs of confidence intervals (Kang and Schmeiser, 1990). Points inside the 45-degree angle with the vertex at θ represent intervals that cover θ. There are 100 confidence intervals of each type in a plot, so coverage can be assessed by counting the number of points outside the lines. Points that are lower (shorter halfwidth) and centered within the 45-degree angle (midpoint close to θ) are preferred. In Figures 4 and 5 the circles are $\hat{\theta}_L \pm H_L$ intervals, and the pluses are a remedy.

Figure 4 is based on the same multivariate normal data as Figure 1a (sample size $n = 10$). The two plots are $\hat{\theta}_L$ versus $\hat{\theta}_J$ and $\hat{\theta}_S$, respectively. The outstanding

Table II
Results for Selected Experiments Based on
$m = 100$ Macroreplications

Measure	Crude	Linear	Batch	Known	Jackknife	Splitting
Multivariate Normal (a), $n = 10$						
MSE	0.130	0.043	0.043	0.072	0.141	0.066
Variance	0.130	0.042	0.042	0.072	0.141	0.066
Bias	−0.008	0.012	0.014	−0.010	−0.019	−0.002
Average Variance[a]	0.097	0.037	0.037	0.076	0.149	0.046
Halfwidth	0.687	0.484	0.485	0.752	0.707	0.440
Coverage	0.92	0.96	0.96	0.97	0.96	0.88
Multivariate Normal (a), $n = 60$						
MSE	0.0171	0.0016	0.0020	0.0039	0.0016	0.0016
Variance	0.0171	0.0016	0.0020	0.0039	0.0016	0.0016
Bias	−0.0021	−0.0007	0.0007	−0.0003	−0.0005	−0.0006
Average Variance	0.0171	0.0022	0.0025	0.0037	0.0024	0.0022
Halfwidth	0.2605	0.0937	0.1028	0.1216	0.0969	0.0930
Coverage	0.98	0.98	1.00	0.97	0.99	0.98
Pearson VII (a), $n = 10$						
MSE	0.090	0.023	0.023	1.898	0.051	0.044
Variance	0.086	0.023	0.023	1.894	0.051	0.044
Bias	0.067	0.011	0.011	0.067	−0.007	0.009
Average Variance	0.090	0.020	0.020	0.067	0.064	0.043
Halfwidth	0.646	0.344	0.344	0.686	0.480	0.398
Coverage	0.94	0.92	0.92	0.91	0.95	0.88
Pearson VII (a), $n = 60$						
MSE	0.0189	0.0021	0.0022	0.0152	0.0025	0.0024
Variance	0.0188	0.0021	0.0022	0.0151	0.0024	0.0024
Bias	0.0097	0.0023	0.0004	0.0112	0.0031	0.0031
Average Variance	0.0166	0.0019	0.0021	0.0034	0.0022	0.0022
Halfwidth	0.2529	0.0861	0.0930	0.1154	0.0932	0.0934
Coverage	0.93	0.92	0.96	0.87	0.91	0.91
Extended KL (a), $n = 10$						
MSE	7.6	3.8	3.8	4.9	9.1	5.2
Variance	7.6	2.7	2.7	4.2	9.1	5.1
Bias	0.2	−1.1	−1.1	−0.8	−0.1	0.2
Average Variance	6.9	3.2	3.2	5.7	12.0	5.0
Halfwidth	5.7	4.5	4.5	6.4	6.8	4.5
Coverage	0.95	0.95	0.95	0.99	0.96	0.94
Extended KL (a), $n = 60$						
MSE	1.02	0.26	0.31	0.29	0.25	0.24
Variance	1.00	0.23	0.28	0.27	0.25	0.24
Bias	0.13	−0.17	−0.15	−0.13	0.00	0.00
Average Variance	1.20	0.26	0.28	0.36	0.32	0.29
Halfwidth	2.18	1.02	1.08	1.19	1.12	1.06
Coverage	0.99	0.96	0.95	0.95	0.98	0.97
M/M/1, $n = 10$						
MSE	0.562	0.480	0.480	0.681	0.800	0.574
Variance	0.562	0.411	0.411	0.613	0.792	0.571
Bias	−0.011	−0.262	−0.262	−0.220	−0.089	−0.059
Average Variance	0.542	0.440	0.440	0.446	1.146	0.494
Halfwidth	1.584	1.505	1.505	1.651	2.123	1.458
Coverage	0.91	0.93	0.93	0.89	0.97	0.90
M/M/1, $n = 60$						
MSE	0.1089	0.0369	0.0375	0.0731	0.0380	0.0379
Variance	0.1070	0.0369	0.0375	0.0728	0.0376	0.0374
Bias	0.0426	−0.0021	−0.0040	−0.0167	0.0206	0.0226
Average Variance	0.0982	0.0315	0.0331	0.0369	0.0363	0.0328
Halfwidth	0.6216	0.3535	0.3700	0.3825	0.3777	0.3596
Coverage	0.92	0.92	0.93	0.88	0.93	0.91

[a] Average Variance is the average of 100 variance estimates and should be compared to Variance.

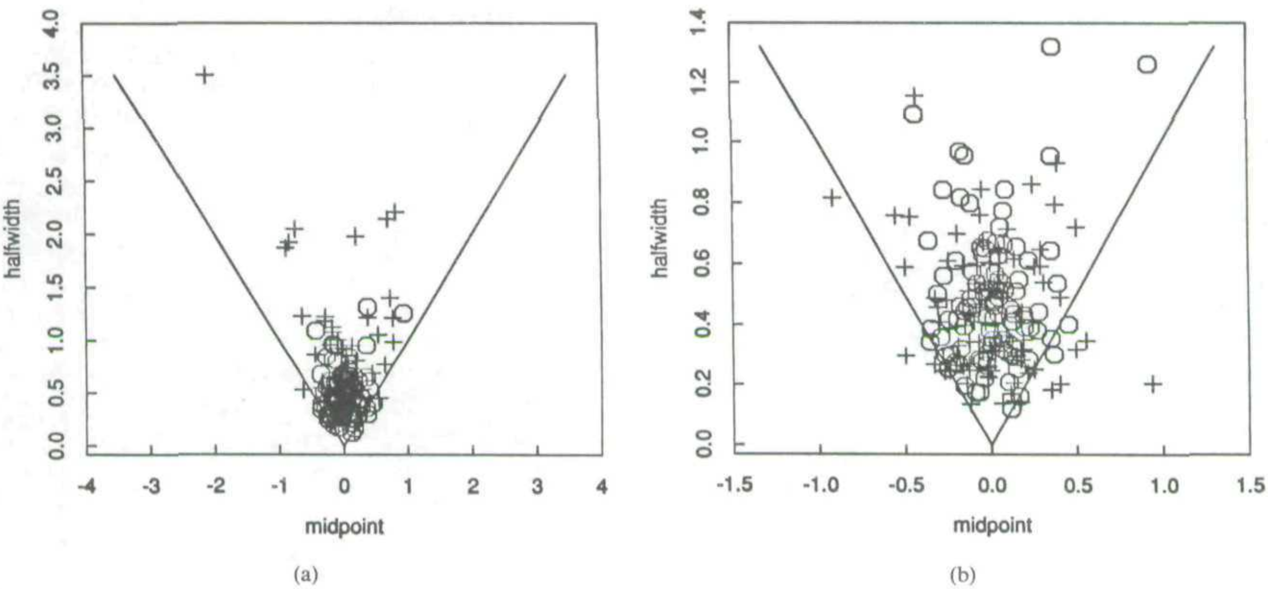


Figure 4. Confidence interval comparison. Midpoint by halfwidth plots for multivariate normal data with $\mathbf{\Sigma}^{a}_{ZZ}$ and $n = 10$. Circles for $\hat{\theta}_L$, versus pluses for a) $\hat{\theta}_J$ and b) $\hat{\theta}_S$.

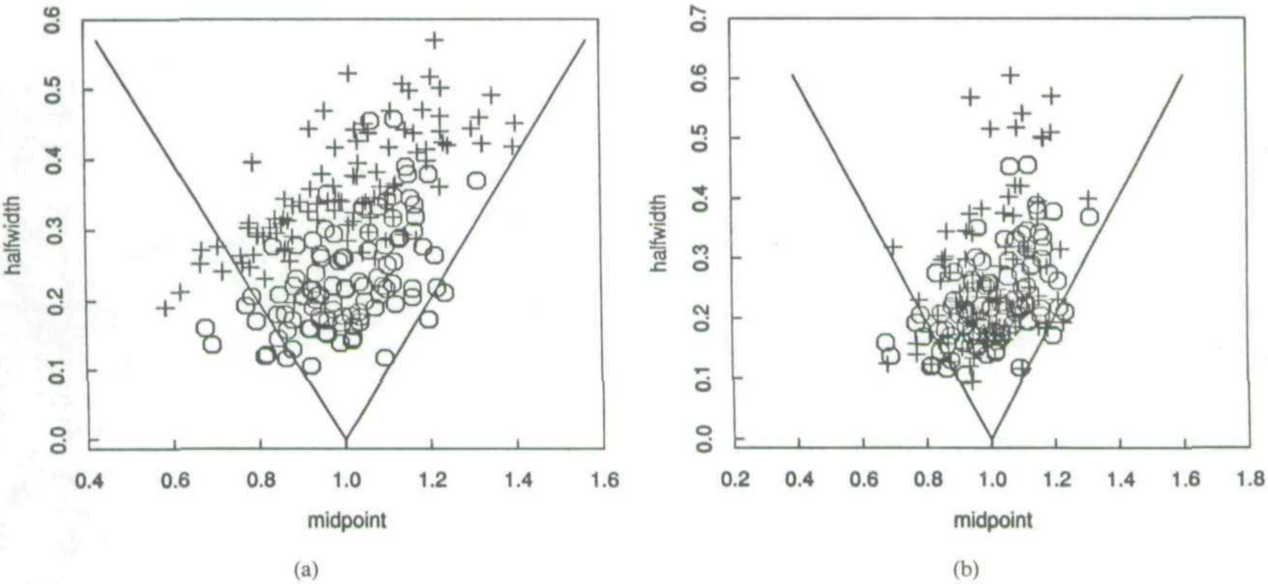


Figure 5. Confidence interval comparison. Midpoint by halfwidth plots for Plackett distribution data with exponential marginals and $n = 30$. Circles for $\hat{\theta}_L$, versus pluses for a) $\hat{\theta}_C$ and b) $\hat{\theta}_B(10)$.

features are the excessive halfwidth for the jackknife estimator and the undercoverage for splitting. This was typical behavior for many data sets at the small sample size.

Figure 5 is based on the bivariate Plackett distribution with exponential marginals when $n = 30$. The two plots are $\hat{\theta}_L$ versus $\hat{\theta}_C$ and $\hat{\theta}_B(10)$, respectively. Positive dependence between mean and variance estimators is apparent in Figure 5a because the intervals that fail to cover θ are too short. Batching appears to improve coverage somewhat. In this example all the intervals appear to have acceptable coverage.

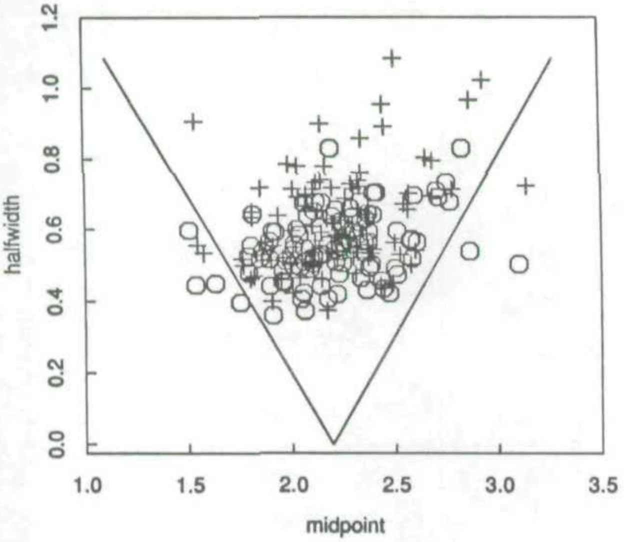


Figure 6. Confidence interval comparison. Midpoint by halfwidth plots for M/M/1 data with $n = 30$. Circles for $\hat{\theta}_S$ versus pluses for $\hat{\theta}_J$.

Figure 6 shows a plot of $\hat{\theta}_J$ versus $\hat{\theta}_S$ for the M/M/1 data when $n = 30$, with the circles representing $\hat{\theta}_S \pm H_S$. The interval based on splitting appears to have nearly correct coverage with shorter halfwidth; this was typical of many examples when $n = 30$. When $n = 60$, all the intervals had acceptable coverage in all the examples.

8.3.4. Bootstrap Estimators

We computed $\hat{\theta}_E$ and the two associated confidence intervals for a very limited set of experiments on multivariate normal and Pearson Type VII data. In these experiments the bootstrap estimator seemed to perform well when n was as large as 30, comparable to $\hat{\theta}_S$. When $n = 10$, nearly singular $\mathbf{X}'\mathbf{X}$ caused erratic behavior. No general conclusions should be drawn from these experiments.

9. RECOMMENDATIONS

Based on the theoretical and empirical results of this study we make the following tentative recommendations based only on sample size n and assuming $q \leq 5$ controls. Certainly there are extreme situations for which these recommendations should be modified, but they provide reasonable guidance when there is no other information available.

1. If $n > 60$, batch the data as close to $k = 60$ batches as possible, then use $\hat{\theta}_B(k)$ and its associated variance estimator and confidence interval. We saw no situation in which the standard linear CV did not work well when the sample size was 60.
2. If n is near 30, use $\hat{\theta}_S$ and its associated variance estimator and confidence interval. This point estimator is always unbiased, its variance is nearly the same as $\hat{\theta}_L$ provided that n is not too small, and the variance estimator and confidence interval are reliable when n is not too small.
3. If $n \leq 10$, $\hat{\theta}_L$ is the most stable point estimator, but $\hat{\theta}_S$ also beats $\hat{\theta}_C$ in terms of variance, and it is unbiased. There is an undercoverage problem for $\hat{\theta}_S \pm H_S$, however, due to underestimation of the variance by S_S^2. We recommend $\hat{\theta}_S$ as the point estimator if bias is suspected to be severe, and $\hat{\theta}_L$ otherwise. We recommend none of the confidence intervals.
4. We do not recommend $\hat{\theta}_K$. The estimator $\hat{\theta}_J$ seems to be dominated by $\hat{\theta}_S$, except for low coverage when $n = 10$.

APPENDIX A

Proofs

Proofs of selected theorems and lemmas needed to support the proofs are given in this appendix. Proofs of the other theorems are in Nelson (1988) or the references.

Lemma A1

$$\mathbf{G} = \begin{pmatrix} n^{-1} + (n-1)^{-1}(\overline{\mathbf{C}} - \boldsymbol{\mu})'\mathbf{S}_{\mathbf{CC}}^{-1}(\overline{\mathbf{C}} - \boldsymbol{\mu}), & -(n-1)^{-1}(\overline{\mathbf{C}} - \boldsymbol{\mu})'\mathbf{S}_{\mathbf{CC}}^{-1} \\ -(n-1)^{-1}\mathbf{S}_{\mathbf{CC}}^{-1}(\overline{\mathbf{C}} - \boldsymbol{\mu}), & (n-1)^{-1}\mathbf{S}_{\mathbf{CC}}^{-1} \end{pmatrix}$$

Proof. See Theorem 8.2.1 in Graybill (1969).

Lemma A2. *If* $\mathbf{X}$ *has rank* $q + 1$ *with probability* 1, *then the matrices* $\mathbf{G}$ *and* $\mathbf{XGX}'$ *are positive definite, and* $\mathbf{H}$ *is positive semidefinite with probability* 1.

Proof. The results cited in the proof are from Seber (1977). Since the matrix $\mathbf{X}$ has rank $q + 1$ with probability 1, $\mathbf{X}'\mathbf{X}$ is positive definite (A.4.6), which implies that $\mathbf{G}$ is positive definite (A.4.3), which in turn implies that $\mathbf{XGX}'$ is positive definite (A.4.5).

The matrix $\mathbf{XGX}'$ is idempotent, which implies that $\mathbf{H} = \mathbf{I}_{n\times n} - \mathbf{XGX}'$ is idempotent (A.5.3). Since $\mathbf{H}$ is also symmetric, it is positive semidefinite (A.5.4).

From here on we assume that $\mathbf{X}$ has rank $q + 1$ with probability 1.

Lemma A3

$$\boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}_i = n^{-1} + (n-1)^{-1}(\overline{\mathbf{C}} - \boldsymbol{\mu})'\mathbf{S}_{\mathrm{CC}}^{-1}(\overline{\mathbf{C}} - \mathbf{C}_i)$$

$$\mathrm{E}[\boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}_i] = 1/n$$

$$\mathbf{X}_i'\mathbf{G}\mathbf{X}_j = n^{-1} + (n-1)^{-1}(\mathbf{C}_i - \overline{\mathbf{C}})'\mathbf{S}_{\mathrm{CC}}^{-1}(\mathbf{C}_j - \overline{\mathbf{C}})$$

$$\mathrm{E}[\mathbf{X}_i'\mathbf{G}\mathbf{X}_i] = (q+1)/n.$$

Proof. The expressions for $\boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}_i$ and $\mathbf{X}_i'\mathbf{G}\mathbf{X}_j$ are directly calculated using Lemma A1. Since

$$\sum_{i=1}^{n} \boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}_i = 1$$

and since the $\{\boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}_i\}$ are identically distributed, their common expectation must be $1/n$. Similarly, $\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{G}\mathbf{X}_i = q + 1$, so their common expectation must be $(q+1)/n$.

Lemma A4. *The diagonal elements of* $\mathbf{H}$, $H_{ii} = 1 - \mathbf{X}_i'\mathbf{G}\mathbf{X}_i$, $i = 1, \ldots, n$, *have expectation* $\mathrm{E}[H_{ii}] = 1 - (q+1)/n$. *If conditions* i *and* ii *of Theorem* 2 *hold, then the* $\{H_{ii}\}$ *are strictly between* 0 *and* 1.

Proof. The expectation can be calculated using Lemma A3. Since $\mathbf{G}$ is positive definite (Lemma A2), $\mathbf{X}_i'\mathbf{G}\mathbf{X}_i > 0$, which implies $H_{ii} < 1$. On the other hand, under conditions i and ii, $\sigma^2 H_{ii} = \mathrm{Var}[\hat{\epsilon}_i \mid \mathbf{C}] > 0$, which implies $H_{ii} > 0$.

Lemma A5. *Under conditions* i *and* ii *of Theorem* 2, $\hat{\boldsymbol{\epsilon}}$ *and* $\hat{\boldsymbol{\gamma}}$ *are conditionally uncorrelated given* $\mathbf{C}$.

Proof

$$\begin{aligned}\mathrm{Cov}[\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\gamma}} \mid \mathbf{C}] &= \mathrm{Cov}[\mathbf{H}\mathbf{Y}, \mathbf{G}\mathbf{X}'\mathbf{Y} \mid \mathbf{C}] \\ &= \mathbf{H}\,\mathrm{Cov}[\mathbf{Y} \mid \mathbf{C}]\mathbf{X}\mathbf{G} \\ &= \sigma^2\mathbf{H}\mathbf{X}\mathbf{G} = \mathbf{0}_{n\times q+1}.\end{aligned}$$

Proof of Theorem 6. Since $\hat{\theta}_J = \hat{\theta}_L + ((n-1)/n)\mathbf{V}'\hat{\boldsymbol{\epsilon}}$, and in light of Theorem 3, the result is proved if we show that $\sqrt{n}((n-1)/n)\mathbf{V}'\hat{\boldsymbol{\epsilon}} \xrightarrow{P} 0$.

Recall that $\hat{\epsilon}_i = Y_i - \mathbf{X}_i'\hat{\boldsymbol{\gamma}}$. From Lemmas A1 and A4

$$\begin{aligned}V_i &= \boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}_i/H_{ii} \\ &= (n^{-1} + (n-1)^{-1}(\overline{\mathbf{C}} - \boldsymbol{\mu})'\mathbf{S}_{\mathrm{CC}}^{-1}(\overline{\mathbf{C}} - \mathbf{C}_i)) \\ &\quad \times (1 - n^{-1} - (n-1)^{-1} \times (\mathbf{C}_i - \overline{\mathbf{C}})'\mathbf{S}_{\mathrm{CC}}^{-1}(\mathbf{C}_i - \overline{\mathbf{C}}))^{-1}.\end{aligned}$$

We need the following intermediate results

$$\hat{\epsilon}_i \Rightarrow Y_i - \mathbf{X}_i'\boldsymbol{\gamma}$$

$$\mathbf{C}_i - \overline{\mathbf{C}} \Rightarrow \mathbf{C}_i - \boldsymbol{\mu}$$

$$(\mathbf{C}_i - \overline{\mathbf{C}})'\mathbf{S}_{\mathrm{CC}}^{-1}(\mathbf{C}_i - \overline{\mathbf{C}}) \Rightarrow (\mathbf{C}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}_{\mathrm{CC}}^{-1}(\mathbf{C}_i - \boldsymbol{\mu})$$

$$(\overline{\mathbf{C}} - \boldsymbol{\mu}) \xrightarrow{P} \mathbf{0}_{q\times 1}.$$

Repeated applications of Slutsky's theorem, and the fact that weak convergence to a constant implies convergence in probability, gives $1/H_{ii} \xrightarrow{P} 1$ and $\sqrt{n}\boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}_i \xrightarrow{P} 0$. Thus, $\sqrt{n}((n-1)/n)\mathbf{V}'\hat{\boldsymbol{\epsilon}} \xrightarrow{P} 0$.

Proof of Theorem 7. Under condition i, $\mathrm{E}[\hat{\theta}_J \mid \mathbf{C}] = \theta + ((n-1)/n)\mathrm{E}[\mathbf{V}'\hat{\boldsymbol{\epsilon}} \mid \mathbf{C}]$. But $\mathrm{E}[\mathbf{V}'\hat{\boldsymbol{\epsilon}} \mid \mathbf{C}] = \mathbf{V}'\mathrm{E}[\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\gamma}} \mid \mathbf{C}] = 0$, which proves unbiasedness. If Equation (1) pertains, then the result is immediate after noticing that

$$\begin{aligned}\mathrm{E}[\tilde{\theta}_i] &= n\left(\eta + \frac{1}{2}\left(\frac{n-2}{2}\right)\mathrm{trace}[\boldsymbol{\Delta}\boldsymbol{\Sigma}_{\mathrm{CC}}]\right) \\ &\quad - (n-1)\left(\eta + \frac{1}{2}\left(\frac{n-3}{2}\right)\mathrm{trace}[\boldsymbol{\Delta}\boldsymbol{\Sigma}_{\mathrm{CC}}]\right) \\ &= \eta + \frac{1}{2}\,\mathrm{trace}[\boldsymbol{\Delta}\boldsymbol{\Sigma}_{\mathrm{CC}}].\end{aligned}$$

Suppose that conditions i and ii hold. Then

$$\begin{aligned}\mathrm{Var}[\hat{\theta}_J] &= \mathrm{Var}[\mathrm{E}[\hat{\theta}_J \mid \mathbf{C}]] + \mathrm{E}[\mathrm{Var}[\hat{\theta}_J \mid \mathbf{C}]] \\ &= \mathrm{E}[\mathrm{Var}[\hat{\theta}_J \mid \mathbf{C}]].\end{aligned}$$

But

$$\mathrm{Var}[\hat{\theta}_J \mid \mathbf{C}] = \mathrm{Var}[\hat{\theta}_L \mid \mathbf{C}] + \left(\frac{n-1}{n}\right)^2 \mathrm{Var}[\mathbf{V}'\mathbf{H}\mathbf{Y} \mid \mathbf{C}]$$

since $\hat{\theta}_L$ and $\hat{\boldsymbol{\epsilon}} = \mathbf{H}\mathbf{Y}$ are conditionally uncorrelated (Lemma A5). Finally

$$\begin{aligned}\mathrm{Var}[\mathbf{V}'\mathbf{H}\mathbf{Y} \mid \mathbf{C}] &= \mathbf{V}'\mathbf{H}\,\mathrm{Var}[\mathbf{Y} \mid \mathbf{C}]\mathbf{H}'\mathbf{V} \\ &= \mathbf{V}'\mathbf{H}(\sigma^2\mathbf{I})\mathbf{H}'\mathbf{V} \\ &= \sigma^2\mathbf{V}'\mathbf{H}\mathbf{V} \geqslant 0\end{aligned}$$

because $\mathbf{H}$ is symmetric, idempotent, and positive semidefinite (Lemma A2). Thus

$$\begin{aligned}\mathrm{Var}[\hat{\theta}_J] &= \mathrm{Var}[\hat{\theta}_L] \\ &\quad + \left(\frac{n-1}{n}\right)^2 \sigma^2\mathrm{E}[\mathbf{V}'\mathbf{H}\mathbf{V}] \geqslant \mathrm{Var}[\hat{\theta}_L].\end{aligned}$$

Proof of Theorem 8. Lemma A4 implies that H_{ii} can be expanded in a power series. Thus, $V_i = \boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}_i/\mathrm{H}_{ii} \approx \boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}_i(1 + \mathbf{X}_i'\mathbf{G}\mathbf{X}_i)$, except for terms of $O(n^{-2})$

in expectation, from Lemma A3. Let $\mathbf{D}$ be the $n \times n$ matrix with diagonal elements $D_{ii} = \mathbf{X}_i' \mathbf{G} \mathbf{X}_i$, and off-diagonal elements 0. Then $\mathbf{V} \approx (\mathbf{I}_{n \times n} + \mathbf{D})\mathbf{X}\mathbf{G}\boldsymbol{\delta}_1$, which in turn implies that

$$\begin{aligned}\mathbf{V}'\mathbf{H}\mathbf{V} &\approx \boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}'(\mathbf{I}_{n\times n} + \mathbf{D})\mathbf{H}(\mathbf{I}_{n\times n} + \mathbf{D})\mathbf{X}\mathbf{G}\boldsymbol{\delta}_1 \\ &= \boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}'\mathbf{D}(\mathbf{I}_{n\times n} - \mathbf{X}\mathbf{G}\mathbf{X}')\mathbf{D}\mathbf{X}\mathbf{G}\boldsymbol{\delta}_1 \\ &\leqslant \boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}'\mathbf{D}^2\mathbf{X}\mathbf{G}\boldsymbol{\delta}_1 \\ &= \sum_{i=1}^{n} (\boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}_i)^2(\mathbf{X}_i'\mathbf{G}\mathbf{X}_i)^2\end{aligned}$$

where the inequality holds because both $\mathbf{I}_{n\times n}$ and $\mathbf{X}\mathbf{G}\mathbf{X}'$ are positive definite (Lemma A2). In expectation, the sum is $nO(n^{-2})O(n^{-2}) = O(n^{-3})$ from Lemma A3.

The proof of Theorem 9 is analogous to the proof of Theorem 6. The proof of Theorem 10 follows the same steps as the proof of Theorem 7.

Proof of Theorem 11. The proof is similar to the proof of Theorem 8. The random variable $W_i \approx (\mathbf{X}_i - \boldsymbol{\delta}_1)'\mathbf{G}\mathbf{X}_i$, up to the terms of $O(n^{-2})$ in expectation. Thus, $\mathbf{W} \approx \mathbf{D}\mathbf{1}_{n\times 1} - \mathbf{X}\mathbf{G}\boldsymbol{\delta}_1$, and

$$\begin{aligned}\mathbf{W}'\mathbf{H}\mathbf{W} &\approx (\mathbf{D}\mathbf{1}_{n\times 1} - \mathbf{X}\mathbf{G}\boldsymbol{\delta}_1)'\mathbf{H}(\mathbf{D}\mathbf{1}_{n\times 1} - \mathbf{X}\mathbf{G}\boldsymbol{\delta}_1) \\ &= \mathbf{1}_{1\times n}\mathbf{D}^2\mathbf{1}_{n\times 1} - \mathbf{1}_{1\times n}\mathbf{D}\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{D}\mathbf{1}_{n\times 1} \\ &\leqslant \mathbf{1}_{1\times n}\mathbf{D}^2\mathbf{1}_{n\times 1} \\ &= \sum_{i=1}^{n} (\mathbf{X}_i'\mathbf{G}\mathbf{X}_i)^2.\end{aligned}$$

The last summation is $nO(n^{-2}) = O(n^{-1})$ in expectation. Thus, $n^{-2}\mathrm{E}[\mathbf{W}'\mathbf{G}\mathbf{W}] = O(n^{-3})$.

Lemma A6

$$n(n-1)S_S^2 = \sum_{i=1}^{n} (1 + W_i)^2 \hat{\epsilon}_i^2 - \frac{1}{n}\mathbf{W}'\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'\mathbf{W}.$$

Proof. From the results in Appendix B we can write

$$\begin{aligned}\overline{\theta}_i - \hat{\theta}_S &= Y_i - (\mathbf{C}_i - \boldsymbol{\mu})'\hat{\boldsymbol{\beta}}^{-i} - \hat{\theta}_S \\ &= Y_i - \mathbf{X}_i'\hat{\boldsymbol{\gamma}}^{-i} + \hat{\theta}_L^{-i} - \hat{\theta}_S \\ &= Y_i - \mathbf{X}_i'\left(\hat{\boldsymbol{\gamma}} - \frac{\mathbf{G}\mathbf{X}_i\hat{\epsilon}_i}{H_{ii}}\right) + \hat{\theta}_L \\ &\qquad - \frac{\boldsymbol{\delta}_1'\mathbf{G}\mathbf{X}_i\hat{\epsilon}_i}{H_{ii}} - \hat{\theta}_L - \frac{\mathbf{W}'\hat{\boldsymbol{\epsilon}}}{n} \\ &= Y_i - \mathbf{X}_i'\hat{\boldsymbol{\gamma}} + W_i\hat{\epsilon}_i - \frac{\mathbf{W}'\hat{\boldsymbol{\epsilon}}}{n} \\ &= (1 + W_i)\hat{\epsilon}_i - \frac{\mathbf{W}'\hat{\boldsymbol{\epsilon}}}{n}.\end{aligned}$$

Thus

$$\begin{aligned}&\sum_{i=1}^{n} (\overline{\theta}_i - \hat{\theta}_S)^2 \\ &\quad = \sum_{i=1}^{n}\left((1 + W_i)^2\hat{\epsilon}_i^2 - \frac{2(1 + W_i)\hat{\epsilon}_i\mathbf{W}'\hat{\boldsymbol{\epsilon}}}{n} + \left(\frac{\mathbf{W}'\hat{\boldsymbol{\epsilon}}}{n}\right)^2\right) \\ &\quad = \sum_{i=1}^{n} (1 + W_i)^2\hat{\epsilon}_i^2 - \frac{\mathbf{W}'\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'\mathbf{W}}{n}\end{aligned}$$

using the fact that $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$.

Proof of Theorem 12. We first prove that $nS_S^2 \xrightarrow{P} \sigma^2$. From Lemma A6 we can write

$$nS_S^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} (1 + 2W_i + W_i^2)\hat{\epsilon}_i^2 - \frac{1}{n}\mathbf{W}'\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'\mathbf{W}\right). \tag{4}$$

Also, $(n-1)^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = n(n-q-1)/(n-1)S^2 \xrightarrow{P} \sigma^2 = \sigma_Y^2(1 - R^2)$ (Nelson 1988). The proof is completed by noticing that $W_i \xrightarrow{P} 0$, which implies that the other terms in (4) go to 0 in probability.

To prove the small-sample result we need some intermediate results. First

$$\begin{aligned}&\mathrm{E}\left[\sum_{i=1}^{n} (1 + 2W_i)\hat{\epsilon}_i^2 \mid \mathbf{C}\right] \\ &\quad = \sigma^2 \sum_{i=1}^{n} (1 + 2W_i)H_{ii} \\ &\quad = \sigma^2\left(\sum_{i=1}^{n} H_{ii} + 2\sum_{i=1}^{n} (\mathbf{X}_i - \boldsymbol{\delta}_1)'\mathbf{G}\mathbf{X}_i\right) \\ &\quad = \sigma^2(n - q - 1 + 2(q + 1 - 1)) \\ &\quad = \sigma^2(n + q - 1)\end{aligned}$$

where the last equality comes from the proof of Lemma A3. Using the fact that $\hat{\boldsymbol{\epsilon}} = \mathbf{H}\mathbf{Y}$ we can show that

$$\mathrm{E}[\mathbf{W}'\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'\mathbf{W} \mid \mathbf{C}] = \sigma^2\mathbf{W}'\mathbf{H}\mathbf{W}.$$

And finally, we can write $\sum_{i=1}^{n} W_i^2\hat{\epsilon}_i^2 = \mathbf{W}'\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'\mathbf{W} - 2\sum_{i\neq j} W_i\hat{\epsilon}_i\hat{\epsilon}_j W_j$, so that

$$\mathrm{E}\left[\sum_{i=1}^{n} W_i^2\hat{\epsilon}_i^2 \mid \mathbf{C}\right] = \sigma^2\mathbf{W}'\mathbf{H}\mathbf{W} - 2\sum_{i\neq j} W_i H_{ij} W_j.$$

Thus

$$\mathrm{E}[S_S^2 \mid \mathbf{C}]$$

$$= \left(\frac{n+q-1}{n-1}\right)\frac{\sigma^2}{n} + \frac{1}{n(n-1)}\,\mathbf{W}'\mathbf{H}\mathbf{W}$$

$$- \frac{2}{n(n-1)} \sum_{i \neq j} W_i H_{ij} W_j - \frac{1}{n^2(n-1)}\,\mathbf{W}'\mathbf{H}\mathbf{W}.$$

Collecting terms and taking the expected value completes the proof.

APPENDIX B

Derivations

The computational expressions for $\hat{\theta}_J$ and $\hat{\theta}_S$, and Equations (2) and (3), are derived in this appendix. From Belsley, Kuh and Welsch (1980, Chapter 2), $\hat{\boldsymbol{\gamma}}^{-i} = \hat{\boldsymbol{\gamma}} - \mathbf{G}\mathbf{X}_i \hat{\epsilon}_i / H_{ii}$, where $\hat{\epsilon}_i = Y_i - \mathbf{X}_i' \hat{\boldsymbol{\gamma}}$ and H_{ii} is the iith element of $\mathbf{H} = \mathbf{I}_{n \times n} - \mathbf{X}\mathbf{G}\mathbf{X}'$. If we let $\hat{\theta}_L^{-i}$ be the first element of $\hat{\boldsymbol{\gamma}}^{-i}$, then $\hat{\theta}_L^{-i} = \hat{\theta}_L - \boldsymbol{\delta}_1' \mathbf{G}\mathbf{X}_i \hat{\epsilon}_i / H_{ii}$. Substituting this expression into

$$\hat{\theta}_J = n^{-1} \sum_{i=1}^{n} (n\hat{\theta}_L - (n-1)\hat{\theta}_L^{-i})$$

collecting terms, and noticing that $\hat{\boldsymbol{\epsilon}} = \mathbf{H}\mathbf{Y}$, yields

$$\hat{\theta}_J = \hat{\theta}_L + \left(\frac{n-1}{n}\right)\mathbf{V}'\mathbf{H}\mathbf{Y}.$$

The expression for $\hat{\theta}_S$ is derived similarly by writing

$$\hat{\theta}_S = n^{-1} \sum_{i=1}^{n} (Y_i - (\mathbf{C}_i - \boldsymbol{\mu})' \hat{\boldsymbol{\beta}}^{-i})$$

$$= n^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i' \hat{\boldsymbol{\gamma}}^{-i} + \hat{\theta}_L^{-i})$$

then substituting the expressions above for $\hat{\boldsymbol{\gamma}}^{-i}$ and $\hat{\theta}_L^{-i}$ and collecting terms.

Suppose that $q = 1$ and condition i of Theorem 2 holds, but $\mathrm{Var}[Y_i \mid C_i] = \nu C_i$, where ν is a constant. Let $\boldsymbol{\Lambda}$ be the $n \times n$ matrix with diagonal elements νC_i, and off-diagonal elements 0; that is, $\boldsymbol{\Lambda} = \mathrm{Var}[\mathbf{Y} \mid \mathbf{C}]$. Notice that $\hat{\theta}_L = (n^{-1}\mathbf{1}_{1\times n} - (\overline{\mathbf{C}} - \boldsymbol{\mu})'(\mathbf{L}'\mathbf{L})^{-1}\mathbf{L}')\mathbf{Y} = \mathbf{K}'\mathbf{Y}$, where $\mathbf{1}_{1\times n}$ is a $1 \times n$ vector of 1s and $\mathbf{L}' = ((\mathbf{C}_1 - \overline{\mathbf{C}}), \ldots, (\mathbf{C}_n - \overline{\mathbf{C}}))$. Then

$$\mathrm{Var}[\hat{\theta}_L] = \mathrm{E}[\mathrm{Var}[\hat{\theta}_L \mid \mathbf{C}]] + \mathrm{Var}[\mathrm{E}[\hat{\theta}_L \mid \mathbf{C}]]$$

$$= \mathrm{E}[\mathrm{Var}[\mathbf{K}'\mathbf{Y} \mid \mathbf{C}]]$$

because condition i implies that the second term on the right-hand side is zero because $\mathrm{E}[\hat{\theta}_L \mid \mathbf{C}]$ is a constant. Now

$$\mathrm{Var}[\mathbf{K}'\mathbf{Y} \mid \mathbf{C}]$$

$$= \mathbf{K}' \mathrm{Var}[\mathbf{Y} \mid \mathbf{C}]\mathbf{K}$$

$$= \frac{\upsilon\overline{C}}{n} - \frac{\upsilon(\overline{C} - \mu)}{2n} + \frac{(\overline{C} - \mu)^2 \sum_{i=1}^{n} (C_i - \overline{C})^2 \upsilon C_i}{(n-1)^2 S_{CC}^2}$$

where we use the fact that $\mathbf{L}'\boldsymbol{\Lambda}\mathbf{1}_{n\times 1} = (n-1)S_{CC}\upsilon$, and $\mathbf{L}'\boldsymbol{\Lambda}\mathbf{L} = \sum_{i=1}^{n} (C_i - \overline{C})^2 \upsilon C_i$. Thus

$$\mathrm{Var}[\hat{\theta}_L] = \mathrm{E}\left[\frac{\upsilon\overline{C}}{n}\right]$$

$$+ \mathrm{E}\left[\frac{(\overline{C} - \mu)^2}{S_{CC}} \frac{\upsilon \sum_{i=1}^{n} (C_i - \overline{C})^2 C_i}{(n-1)^2 S_{CC}}\right].$$

On the other hand, $\mathrm{E}[S_L^2] = \mathrm{E}[\mathrm{E}[S^2 G_{11} \mid \mathbf{C}]] = \mathrm{E}[\mathrm{E}[S^2 \mid \mathbf{C}]G_{11}]$, and

$$(n-2)\mathrm{E}[S^2 \mid \mathbf{C}] = \mathrm{E}[\mathbf{Y}'\mathbf{H}\mathbf{Y} \mid \mathbf{C}]$$

$$= \mathrm{trace}[\mathbf{H}\boldsymbol{\Lambda}] + (\mathbf{X}\boldsymbol{\gamma})'\mathbf{H}(\mathbf{X}\boldsymbol{\gamma})$$

$$= \mathrm{trace}[\mathbf{H}\boldsymbol{\Lambda}].$$

Using Lemmas A1 and A3

$$\mathrm{trace}[\mathbf{H}\boldsymbol{\Lambda}]G_{11}$$

$$= \sum_{i=1}^{n} \left(\left(1 - n^{-1} - \frac{(n-1)^{-1}(C_i - \overline{C})^2}{S_{CC}}\right)\nu C_i\right) G_{11}$$

$$= \sum_{i=1}^{n} \left((n-1)\nu\overline{C} - \frac{(n-1)^{-1}\sum_{i=1}^{n} (C_i - \overline{C})^2 \nu C_i}{S_{CC}}\right) G_{11}$$

$$= \left(\frac{n-1}{n} + \frac{(\overline{C} - \mu)^2}{S_{CC}}\right)\nu\overline{C}$$

$$- \left(\frac{n-1}{n} + \frac{(\overline{C} - \mu)^2}{S_{CC}}\right)\frac{\nu \sum_{i=1}^{n} (C_i - \overline{C})^2 C_i}{(n-1)^2 S_{CC}}.$$

Thus

$$\mathrm{E}[S_L^2] = \mathrm{E}\left[\left(\frac{n-1}{n} + \frac{(\overline{C} - \mu)^2}{S_{CC}}\right)\frac{\nu\overline{C}}{n-2}\right] - \frac{1}{n-2}$$

$$\cdot\, \mathrm{E}\left[\left(\frac{n-1}{n} + \frac{(\overline{C} - \mu)^2}{S_{CC}}\right)\frac{\nu \sum_{i=1}^{n} (C_i - \overline{C})^2 C_i}{(n-1)^2 S_{CC}}\right].$$

ACKNOWLEDGMENT

The author acknowledges the helpful comments of the Area Editor, Associate Editor, and two referees. The research was partially supported by National Science Foundation grant ESC-8707634.

REFERENCES

AÑONUEVO, R., AND B. L. NELSON. 1988. Automated Estimation and Variance Reduction via Control

Variates for Infinite-Horizon Simulations. *Comput. Opns. Res.* **15,** 447–456.

ARVENSEN, J. N. 1969. Jackknifing U-Statistics. *Ann. Math. Stat.* **40,** 2076–2100.

BAUER, K. W. 1987. Control Variate Selection for Multiresponse Simulation. Ph.D. Dissertation. School of Industrial Engineering, Purdue University, W. Lafayette, Ind.

BAUER, K. W., S. VENKATRAMAN AND J. R. WILSON. 1987. Estimation Procedures Based on Control Variates With Known Covariance Matrix. In *Proceedings of the Winter Simulation Conference,* 334–341.

BEALE, E. M. L. 1985. Regression: A Bridge Between Analysis and Simulation. *The Statistician* **34,** 141–154.

BECKER, R. A., AND J. M. CHAMBERS. 1984. *S—An Interactive Environment for Data Analysis and Graphics.* Wadsworth, Belmont, Calif.

BELSLEY, D. A., E. KUH AND R. E. WELSCH. 1980. *Regression Diagnostics.* John Wiley, New York.

BRATLEY, P., B. L. FOX AND L. E. SCHRAGE. 1987. *A Guide to Simulation.* Springer-Verlag, New York.

CHENG, R. C. H. 1978. Analysis of Simulation Experiments Under Normality Assumptions. *J. Opnl. Res. Soc.* **29,** 493–497.

EFRON, B. 1982. The Jackknife, the Bootstrap and Other Resampling Plans. *SIAM CBMS-National Science Foundation Monograph* **38.**

EFRON, B., AND G. GONG. 1983. A Leisurely Look at the Bootstrap, the Jackknife, and Cross Validation. *Amer. Statistician* **37,** 36–48.

EFRON, B., AND C. STEIN. 1981. The Jackknife Estimate of Variance. *Ann. Stat.* **9,** 586–596.

EFRON, B., AND R. TIBSHIRANI. 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statist. Sci.* **1,** 54–77.

GLYNN, P. W., AND W. WHITT. 1989. Indirect Estimation Via $L = \lambda W$. *Opns. Res.* **37,** 82–103.

GRAYBILL, F. A. 1969. *Introduction to Matrices With Applications in Statistics.* Wadsworth, Belmont, Calif.

HINKLEY, D. V. 1977. Jackknifing in Unbalanced Situations. *Technometrics* **19,** 285–292.

JOHNSON, M. E. 1987. *Multivariate Statistical Simulation.* John Wiley, New York.

KANG, K., AND B. SCHMEISER. 1990. Methods for Evaluating and Comparing Confidence-Interval Procedures. *Opns. Res.* **38,** 546–552.

KOTTAS, J. F., AND H. LAU. 1978. On Handling Dependent Random Variables in Risk Analysis. *J. Opns. Res. Soc.* **29,** 1209–1217.

LAVENBERG, S. S., AND P. D. WELCH. 1981. A Perspective on the Use of Control Variables to Increase the Efficiency of Monte Carlo Simulations. *Mgmt. Sci.* **27,** 322–335.

LAVENBERG, S. S., T. L. MOELLER AND C. H. SAUER. 1979. Concomitant Control Variables Applied to the Regenerative Simulation of Queueing Systems. *Opns. Res.* **27,** 134–160.

LAVENBERG, S. S., T. L. MOELLER AND P. D. WELCH. 1982. Statistical Results on Control Variables With Application to Queueing Network Simulation. *Opns. Res.* **30,** 182–202.

MILLER, R. G. 1974. An Unbalanced Jackknife. *Ann. Stat.* **2,** 880–891.

NELSON, B. L. 1987a. A Perspective on Variance Reduction in Dynamic Simulation Experiments. *Commun. Statist.* **B16,** 385–426.

NELSON, B. L. 1987b. On Control Variate Estimators. *Comput. Opns. Res.* **14,** 218–225.

NELSON, B. L. 1988. Control-Variate Remedies. Working Paper Series Number 1988-004, Dept. of Industrial and Systems Engineering, Ohio State University, Columbus.

NELSON, B. L. 1989. Batch Size Effects on the Efficiency of Control Variates in Simulation. *Eur. J. Opnl. Res.* **43,** 184–196.

NOZARI, A., S. F. ARNOLD AND C. D. PEGDEN. 1984. Control Variates for Multipopulation Simulation Experiments. *IIE Trans.* **16,** 159–169.

PORTA NOVA, A. M. O., AND J. R. WILSON. 1986. Using Control Variates to Estimate Multiresponse Simulation Metamodels. *1986 Winter Simulation Conference Proceedings,* 326–334.

RIPLEY, B. D. 1987. *Stochastic Simulation.* John Wiley, New York.

RUBINSTEIN, R. Y., AND R. MARKUS. 1985. Efficiency of Multivariate Control Variates in Monte Carlo Simulation. *Opns. Res.* **33,** 661–677.

SCHMEISER, B. 1982. Batch Size Effects in the Analysis of Simulation Output. *Opns. Res.* **30,** 556–568.

SEBER, G. A. F. 1977. *Linear Regression Analysis.* John Wiley, New York.

TEW, J. D., AND J. R. WILSON. 1989. Estimating Simulation Metamodels Using Integrated Variance Reduction Techniques. Technical Report SMS 89-16, School of Industrial Engineering, Purdue University, West Lafayette, Ind.

TOCHER, K. D. 1963. *The Art of Simulation.* The English Universities Press, London.

VENKATRAMAN, S., AND J. R. WILSON. 1986. The Efficiency of Control Variates in Multiresponse Simulation. *Opns. Res. Lett.* **5,** 37–42.

WILSON, J. R. 1984. Variance Reduction Techniques for Digital Simulation. *Am. J. Math. Mgmt. Sci.* **4,** 277–312.

WILSON, J. R., AND A. A. B. PRITSKER. 1984a. Variance Reduction in Queueing Simulation Using Generalized Concomitant Variables. *J. Statist. Comput. Simul.* **19,** 129–153.

WILSON, J. R., AND A. A. B. PRITSKER. 1984b. Experimental Evaluation of Variance Reduction Techniques for Queueing Simulation Using Generalized Concomitant Variables. *Mgmt. Sci.* **30,** 1459–1472.