IMPROVED ESTIMATORS OF MEAN RESPONSE IN SIMULATION WHEN CONTROL VARIATES WITH UNKNOWN COVARIANCE MATRIX ARE USED

by

Leon Jay Gleser
Purdue University

Technical Report #87-34

Department of Statistics
Purdue University

July 1987

★

IMPROVED ESTIMATORS OF MEAN RESPONSE IN SIMULATION WHEN CONTROL VARIATES WITH UNKNOWN COVARIANCE MATRIX ARE USED

by

Leon Jay Gleser
Purdue University

Technical Report #87-34

Department of Statistics
Purdue University

July 1987

★

IMPROVED ESTIMATORS OF MEAN RESPONSE IN SIMULATION WHEN CONTROL VARIATES WITH UNKNOWN COVARIANCE MATRIX ARE USED

Leon Jay Gleser[1]
Purdue University

ABSTRACT

In discrete event simulation the method of control variates is often used to reduce the variance of estimation for the mean of the output response. The control variates $x_1, \ldots, x_p$ have known means and are assumed to have a joint normal distribution with the output response y. Consequently, the mean of y is estimated by linear regression.

In the present paper, it is shown that when the covariance matrix of the vector $(x_1, \ldots, x_p)$ of control variables is unknown, and three or more control variables are used, the usual linear regression estimator of the mean of y is one of a large class of unbiased estimators, many of which have smaller variance than the usual estimator. These estimators are adaptive to information in the data concerning the multiple correlation between the dependent variable and the control variables.

Key words: Discrete event simulation, linear regression, adaptive estimators, minimax estimation, reduced variance, adjusted estimation of the mean.

[1]Research supported by NSF Grant DMS 8501966.

IMPROVED ESTIMATORS OF MEAN RESPONSE IN SIMULATION WHEN CONTROL VARIATES WITH UNKNOWN COVARIANCE MATRIX ARE USED

Leon Jay Gleser

Purdue University

1. Introduction

One of the main variance reduction techniques used in discrete event simulation is the method of control variates. This method attempts to exploit correlations between output responses y and certain associated auxiliary variables $x_1, x_2, \ldots, x_p$ observed during the course of each simulation run. The means $\mu_1, \mu_2, \ldots, \mu_p$ of the auxiliary variables are typically known; the goal is to estimate the mean μ_y of y.

The literature on the use of control variables in simulation is fairly new. The first comprehensive discussion appears in Kleijen (1974). More recent surveys are Wilson (1984) and Bauer (1987).

The model underlying the use of control variables is that of linear regression with random predictors. It is assumed that n repetitions of a simulation experiment yield statistically independent observations

$$(y_i, x_{1i}, x_{2i}, \ldots, x_{pi})' \equiv (y_i, \mathbf{x}_i'),\ i = 1, 2, \ldots, n,$$

on the output response y and the vector $\mathbf{x} = (x_1, \ldots, x_p)'$ of auxiliary (control) variables. Since y and $\mathbf{x}$ result from a common set of generated random numbers and a common probabilistic structure (for example, a multiserver queue), these variables have a joint distribution with mean vector

$$\mu = (\mu_y, \mu_{\mathbf{x}}') = (\mu_y, \mu_1, \mu_2, \ldots, \mu_p)'$$

and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{yy} & \sigma_{y\mathbf{x}} \\ \sigma_{y\mathbf{x}}' & \Sigma_{\mathbf{xx}} \end{pmatrix}.$$

The mean vector $\mu_{\mathbf{x}}$ of the control variables is known (usually from theoretical distributional information concerning these variables).

In the literature on the use of control variables, it is usually assumed that y and $\mathbf{x}$ have a joint normal distribution. Consequently, the conditional distribution of y given $\mathbf{x} = \mathbf{x}_0$ is normal with conditional mean

$$\mu_{y|\mathbf{x}=\mathbf{x}_0} = \mu_y + \beta'(\mathbf{x}_0 - \mu_{\mathbf{x}})$$

and conditional variance

$$\sigma_{yy\cdot\mathbf{x}} = \sigma_{yy}(1 - \rho^2_{y\cdot\mathbf{x}}),$$

where

$$\beta = \sigma_{y\mathbf{x}}\Sigma^{-1}_{\mathbf{xx}}, \quad \rho^2_{y\cdot\mathbf{x}} = \frac{\sigma_{y\mathbf{x}}\Sigma^{-1}_{\mathbf{xx}}\sigma'_{y\mathbf{x}}}{\sigma_{yy}},$$

are the vector of slopes for the regression of y on $\mathbf{x}$ and the squared multiple correlation (coefficient of determination), respectively.

Let

$$(\overline{y}, \overline{\mathbf{x}}') = \frac{1}{n}\sum_{i=1}^{n}(y_i, \mathbf{x}'_i)$$

and

$$W = \begin{pmatrix} w_{yy} & w_{y\mathbf{x}} \\ w_{\mathbf{x}y} & W_{\mathbf{xx}} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n}(y_i - \overline{y})^2 & \sum_{i=1}^{n}(y_i - \overline{y})(\mathbf{x}_i - \overline{\mathbf{x}})' \\ \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(y_i - \overline{y}) & \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' \end{pmatrix}$$

be the sample mean vector and sample cross-product matrix.

In the absence of data from the auxiliary variables $x_1, x_2, \ldots, x_p$, the obvious unbiased estimator of μ_y is $\overline{y}$, which has variance $n^{-1}\sigma_{yy}$. If data from the auxiliary variables is available, and the vector β of slopes is known, then

$$\overline{y}(\beta) = \overline{y} - \beta'(\overline{\mathbf{x}} - \mu_{\mathbf{x}}) \tag{1.1}$$

is an unbiased estimator of μ_y with variance

$$\text{var}\ (\overline{y}(\beta)) = n^{-1}\sigma_{yy\cdot\mathbf{x}} = n^{-1}\sigma_{yy}(1 - \rho^2_{y\cdot\mathbf{x}}).$$

The estimator $\overline{y}(\beta)$ has smaller variance than $\overline{y}$ whenever $\rho^2_{y\cdot\mathbf{x}} > 0$.

Of course, β is typically not known. In this case, we can replace β in (1.1) by the usual least squares estimator

$$\hat{\mathbf{b}} = W_{\mathbf{xx}}^{-1} w_{\mathbf{x}y} \tag{1.2}$$

of β. The resulting estimator

$$\overline{y}(\hat{\mathbf{b}}) = \overline{y} - \hat{\mathbf{b}}'(\overline{\mathbf{x}} - \mu_{\mathbf{x}}) \tag{1.3}$$

is the maximum likelihood estimator of μ_y. This estimator is unbiased for μ_y and has variance

$$\text{var}\ (\overline{y}(\hat{\mathbf{b}})) = n^{-1}(1 + \frac{p}{n-p-2})\ \sigma_{yy}(1 - \rho^2_{y\cdot\mathbf{x}}) \tag{1.4}$$

when $n > p+2$. (When $n \leq p+2$, $\overline{y}(\hat{\mathbf{b}})$ has infinite variance.) Consequently, the estimator (1.4) is superior to $\overline{y}$ (has smaller variance) as an estimator of μ_y if and only if

$$n > p + 2 \textit{ and } \rho^2_{y\cdot\mathbf{x}} > \frac{p}{n-2}. \tag{1.5}$$

However, $\overline{y}(\hat{\mathbf{b}})$ is not the only estimator of the form (1.1) that is unbiased for μ_y. In Gleser (1987), it is shown that for any p-dimensional vector-valued function $\mathbf{b} = \mathbf{b}(W)$ of the sample cross-product matrix W for which the expected value $E[\mathbf{b}(W)]$ exists, all Σ, the estimator

$$\overline{y}(\mathbf{b}) = \overline{y} - \mathbf{b}'(\overline{\mathbf{x}} - \mu_{\mathbf{x}}) \tag{1.6}$$

is an unbiased estimator of μ_y with variance

$$\text{var}\ (\overline{y}(\mathbf{b})) = n^{-1}\sigma_{yy}(1 - \rho^2_{y\cdot\mathbf{x}}) \left\{ 1 + \frac{E[(\mathbf{b} - \beta)'\Sigma_{\mathbf{xx}}(\mathbf{b} - \beta)]}{\sigma_{yy\cdot\mathbf{x}}} \right\}. \tag{1.7}$$

Thus, $\overline{y}(\mathbf{b})$ has uniformly smaller variance than $\overline{y}(\hat{\mathbf{b}})$ if and only if $\mathbf{b}$ dominates $\hat{\mathbf{b}}$ in risk as an estimator of β under the loss function

$$L(\mathbf{b}; \beta, \sigma_{yy\cdot\mathbf{x}}, \Sigma_{\mathbf{xx}}) = \frac{(\mathbf{b} - \beta)'\Sigma_{\mathbf{xx}}(\mathbf{b} - \beta)}{\sigma_{yy\cdot\mathbf{x}}}. \tag{1.8}$$

Gleser (1987) considers the special case where the covariance matrix $\Sigma_{\mathbf{xx}}$ of the vector $\mathbf{x}$ of control variables is known. Using and extending results of Berger (1975), a class of estimators of β is obtained, each of which dominates $\hat{\mathbf{b}}$ in risk when $p \geq 3, n > p+2$. Each member of this class thus yields an unbiased estimator $\overline{y}(\mathbf{b})$ of μ_y which has smaller variance than $\overline{y}(\hat{\mathbf{b}})$. The relative (proportional) saving in variance can be substantial, particularly when $\rho^2_{y\cdot\mathbf{x}}$ is small, and p is large.

In the present paper we consider the case where $\Sigma_{\mathbf{xx}}$ is unknown. In this case, the results of Stein (1960) show that there exist estimators $\mathbf{b}$ of β that improve upon $\hat{\mathbf{b}}$ in risk under the loss function (1.8). (Stein assumes that both μ_y and $\mu_{\mathbf{x}}$ are known. Consequently, his n corresponds to our $n-1$.) At the end of his paper, Stein asks whether one can improve upon $\overline{y}(\hat{\mathbf{b}})$ as an estimator of μ_y when μ_y is unknown, but $\mu_{\mathbf{x}}$ is known. From Equation (1.7), it is apparent that Stein's own results provide an affirmative answer to his question. However, Stein's argument is restricted to a very small class of estimators, none of which can be admissible, and he does not provide sufficient conditions for risk domination.

In Section 2, we exhibit a class of estimators of β which is hopefully large enough to include some admissible estimators. Every member of this class is shown to be adaptive to the information in the data concerning the magnitude of $\rho^2_{y\cdot\mathbf{x}}$. Corresponding to each such estimator $\mathbf{b}$ of β, there is a corresponding unbiased estimator $\overline{y}(\mathbf{b})$ of μ_y. Theorem 1 of Section 2 gives sufficient conditions for an estimator $\mathbf{b}$ in the class to dominate $\hat{\mathbf{b}}$ in risk as an estimator of β under the loss function (1.8). These conditions then are also sufficient for $\overline{y}(\mathbf{b})$ to have everywhere smaller variance than $\overline{y}(\hat{\mathbf{b}})$.

The proof of Theorem 1, which is somewhat complicated, is given in Section 3. Readers interested only in applying the results of this paper may wish to skip this section on first reading. Finally, some indication of the relative improvement in variance that can be obtained by use of a particular member of the class of estimators described by Theorem 1 is given in Section 4.

2. A Class of Adaptive Estimators

Let

$$w_{yy\cdot\mathbf{x}} = w_{yy} - w_{y\mathbf{x}} W_{\mathbf{xx}}^{-1} w_{\mathbf{x}y}.$$

Consider the class of estimators

$$\mathbf{b}_h = \left[1 - h\left(\frac{\hat{\mathbf{b}}' W_{\mathbf{xx}} \hat{\mathbf{b}}}{w_{yy\cdot\mathbf{x}}}\right)\right] \hat{\mathbf{b}} \tag{2.1}$$

of β, where $h(u) = r(u)/u$ and the function $r(u)$ satisfies

$$\begin{aligned} &(i)\ r(\cdot) \text{ maps } [0,\infty) \text{ to } [0,\infty), \\ &(ii)\ r(u) \text{ is nondecreasing in } u, u \geq 0. \end{aligned} \tag{2.2}$$

These estimators depend on the data only through $\hat{\mathbf{b}}$ and

$$T = \frac{\hat{\mathbf{b}}' W_{\mathbf{xx}} \hat{\mathbf{b}}}{w_{yy\cdot\mathbf{x}}} = \frac{R^2}{1-R^2},$$

where R^2 is the sample coefficient of determination.

The estimators

$$\left(1 - \frac{a_1(1-R^2)}{a_2(1-R^2)+R^2}\right)\hat{\mathbf{b}}, \quad a_1 \geq 0,\ a_2 \geq 0,$$

considered by Stein (1960) are members of the class (2.1), (2.2) with $h(u) = a_1(a_2+u)^{-1}$. Stein showed that for sufficiently small a_1 and sufficiently large a_2 such estimators dominate $\hat{\mathbf{b}}$ in risk under the loss function (1.8) when $p \geq 3,\ n > p+2$.

For each member $\mathbf{b}_h$ of the class of estimators (2.1), (2.2) of β there is a corresponding estimator

$$\overline{y}(\mathbf{b}_h) = \overline{y} - \mathbf{b}_h'(\overline{\mathbf{x}} - \mu_{\mathbf{x}}) \tag{2.3}$$

of μ_y.

Recall that $T = (1-R^2)^{-1}R^2$ is a test statistic for testing $H_0 : \rho_{y\cdot\mathbf{x}} = 0$. Indeed, T stochastically increasing in $\rho^2_{y\cdot\mathbf{x}}$. As remarked in Section 1, we would use $\overline{y}$ to estimate μ_y

if $\rho^2_{y\cdot\mathbf{x}} = 0$ and otherwise would wish to use $\overline{y}(\beta)$ or, in ignorance of $\beta, \overline{y}(\hat{\mathbf{b}})$. Thus, it is of interest to note that

$$(2.4) \qquad \begin{aligned} \overline{y}(\mathbf{b}_h) &= \overline{y} - \mathbf{b}_h'(\overline{\mathbf{x}} - \mu_{\mathbf{x}}) \\ &= h(T)\overline{y} + (1 - h(T))\overline{y}(\hat{\mathbf{b}}) \end{aligned}$$

is adaptive to the information provided in the data about the magnitude of $\rho^2_{y\cdot\mathbf{x}}$. If $h(u)$ is nonincreasing in u, small values of T lead to greater weight being placed on $\overline{y}$ as an estimator of μ_y, while large values of T favor $\overline{y}(\hat{\mathbf{b}})$. As shown by Sclove, Morris and Radhakrishnan (1972) a smooth adaptation to information in the data about the parameters is usually preferable to using a preliminary test to choose an estimator.

We now present the main theorem of this paper.

THEOREM 1. Let $p \geq 3, n > p + 2$. Let $\mathbf{b}_h$ be any of the estimators defined by (2.1) and (2.2). If

$$(2.5) \qquad r(u) \equiv uh(u) < \frac{2(p-2)}{n-p+1}, \text{ all } u \geq 0,$$

then under the loss function (1.8), $\mathbf{b}_h$ has everywhere (over $\beta, \sigma_{yy\cdot\mathbf{x}}, \Sigma_{\mathbf{xx}}$) smaller risk than $\hat{\mathbf{b}}$. Consequently, $\overline{y}(\mathbf{b}_h)$ is an unbiased estimator of μ_y with variance uniformly smaller than that of $\overline{y}(\hat{\mathbf{b}})$.

3. Proof of Theorem 1

The proof of Theorem 1 proceeds in four main steps: reduction to a canonical distributional form by invariance, integration by parts to produce an unbiased estimator of risk difference, simplification of expected values, and then a final argument based on an integration by parts for discrete variables (Hwang, 1982).

Reduction by Invariance

The following arguments are essentially those used by Stein (1960). Recall that every estimator $\mathbf{b}_h$ in the class (2.1), (2.2) depends on the data only through $\hat{\mathbf{b}}$ and R^2. It is well known that R^2 is invariant under the group of transformations

$$(3.1) \qquad y \to a_1 Y + a_2, \quad \mathbf{x} \to A\mathbf{x} + \mathbf{a},$$

where a_1, a_2 are arbitrary scalars, A is an arbitrary p-dimensional nonsingular matrix, and $\mathbf{a}$ is an arbitrary p-dimensional column vector. Under such transformations

$$\hat{\mathbf{b}} \to a_1(A')^{-1}\hat{\mathbf{b}}, \quad \mathbf{b}_h \to a_1(A')^{-1}\mathbf{b}_h$$

and also

$$\beta \to a_1(A')^{-1}\beta,\ \sigma_{yy\cdot\mathbf{x}} \to a_1^2\sigma_{yy\cdot\mathbf{x}},\ \Sigma_{\mathbf{xx}} \to A\Sigma_{\mathbf{xx}}A'.$$

Further, for $\mathbf{b} = \hat{\mathbf{b}}$ or $\mathbf{b}_h$,

$$L(a_1(A')^{-1}\mathbf{b}; a_1(A')^{-1}\beta,\ a_1^2\sigma_{yy\cdot\mathbf{x}},\ A\Sigma_{\mathbf{xx}}A') = L(\mathbf{b};\beta,\ \sigma_{yy\cdot\mathbf{x}},\ \Sigma_{\mathbf{xx}}).$$

Consequently, (3.1) leaves invariant the problem of determining the risks of $\hat{\mathbf{b}},\ \mathbf{b}_h$.

Choose

$$a_1 = \sigma_{yy\cdot\mathbf{x}}^{-1/2},\ a_2 = -\sigma_{yy\cdot\mathbf{x}}^{-1/2}\mu_y,\ A = \Gamma\Sigma_{\mathbf{xx}}^{-1/2},\ \mathbf{a} = -A\mu_{\mathbf{x}},$$

where $\Sigma_{\mathbf{xx}}^{1/2}$ is the symmetric square root of $\Sigma_{\mathbf{xx}}$ and Γ is a p-dimensional orthogonal matrix with first row

$$(\beta'\Sigma_{\mathbf{xx}}\beta)^{-1/2}\beta'\Sigma_{\mathbf{xx}}^{1/2}.$$

Under this choice of the transformation (3.1), $\mu' = (\mu_y,\ \mu_{\mathbf{x}}')$ is transformed to the zero vector and

$$\Sigma_{\mathbf{xx}} = \begin{pmatrix} \sigma_{yy} & \sigma_{y\mathbf{x}} \\ \sigma_{\mathbf{x}y} & \Sigma_{\mathbf{xx}} \end{pmatrix} \to \begin{pmatrix} 1+\tau & \gamma' \\ \gamma & I_p \end{pmatrix}$$

where

$$\gamma' = \tau^{1/2}(1,0,0,\ldots,0), \quad \tau = \frac{\beta'\Sigma_{\mathbf{xx}}\beta}{\sigma_{yy\cdot\mathbf{x}}} = \frac{\rho_{y\cdot\mathbf{x}}^2}{1-\rho_{y\cdot\mathbf{x}}^2}.$$

Therefore, we may assume without loss of generality that $\sigma_{yy\cdot\mathbf{x}} = 1+\tau-\gamma'\gamma = 1$, and that

$$W = \begin{pmatrix} w_{yy} & w_{y\mathbf{x}} \\ w_{\mathbf{x}y} & W_{\mathbf{xx}} \end{pmatrix} \sim \mathcal{W}_{p+1}\left(m, \begin{pmatrix} 1+\tau & \gamma' \\ \gamma & I_p \end{pmatrix}\right), \tag{3.2}$$

where $m = n-1$ and $\mathcal{W}_s(\nu,\psi)$ denotes the s-dimensional Wishart distribution with ν degrees of freedom and parameter ψ. The risks of $\mathbf{b}_h$ and $\hat{\mathbf{b}}$ under the loss function (1.8) then depend only on m, p and τ. Let

$$\omega(\tau) = E[(\hat{\mathbf{b}}-\gamma)'(\hat{\mathbf{b}}-\gamma)] - E[(\mathbf{b}_h-\gamma)'(\mathbf{b}_h-\gamma)]$$

be the difference in risks between $\hat{\mathbf{b}}$ and $\mathbf{b}_h$. We need to show that $\omega(\tau) \geq 0$ for all $\tau \geq 0$.

To eliminate repetitious subscripts, let

$$\mathbf{z} = \hat{\mathbf{b}},\ V = W_{\mathbf{xx}},\ w = w_{yy\cdot\mathbf{x}}. \tag{3.3}$$

The following distributional results, which are a consequence of (3.2) and (3.3) are well known [see, for example, Muirhead (1982; Chapter 3)]:

$$\begin{aligned} &w \text{ is independent of } (\mathbf{z},\ V) \\ &w \sim \chi^2_{m-p},\ V \sim \mathcal{W}_p(m,\ I_p), \\ &\mathbf{z}|V \sim \text{ MVN } (\gamma,\ V^{-1}). \end{aligned} \tag{3.4}$$

By a standard argument,

$$\omega(\tau) = E\left[2h\left(\frac{\mathbf{z}'V\mathbf{z}}{w}\right)(\mathbf{z}-\gamma)'\mathbf{z} - h^2\left(\frac{\mathbf{z}'V\mathbf{z}}{w}\right)\mathbf{z}'\mathbf{z}\right]. \tag{3.5}$$

Integration by Parts

We will make use of the following two well-known integration-by-parts results.

Lemma 1. Let $\mathbf{z} \sim$ MVN (η, ψ) and let the real-valued function $g(\mathbf{z})$ be sufficiently regular in $\mathbf{z}$ for integration by parts. Let Q be a real p-dimensional symmetric matrix and

$$\nabla g(\mathbf{z}) = \left(\frac{\partial g(\mathbf{z})}{\partial z_1}, \ldots, \frac{\partial g(\mathbf{z})}{\partial z_p}\right)'.$$

Then if $E[g(\mathbf{z})\mathbf{z}'\mathbf{z}]$ exists,

$$E[g(\mathbf{z})(\mathbf{z}-\eta)'\mathbf{z}] = E[g(\mathbf{z})\ tr(Q\psi) + \mathbf{z}'Q\psi \nabla g(\mathbf{z})].$$

Lemma 2. Let $w \sim \theta^2\chi^2_\nu$. If $g(w)$ is sufficiently regular in w for integration by parts, and if $E[wg(w)]$ exists, then

$$E\left[wg(w)\right] = 2\theta^2 E\left[\frac{\nu}{2}\ g(w) + wg^{(1)}(w)\right],$$

where

$$g^{(1)}(w) = \frac{d}{dw}\, g(w).$$

Taking expected values in the order $E_{w,V}E_{\mathbf{z}|V}$ and using Lemma 1 to integrate by parts over $\mathbf{z}$ yields

$$(3.6) \quad E[h(\frac{\mathbf{z}'V\mathbf{z}}{w})(\mathbf{z}-\gamma)'\mathbf{z}] = E\left[h\left(\frac{\mathbf{z}'V\mathbf{z}}{w}\right)\left(tr\ V^{-1} - 2\frac{\mathbf{z}'\mathbf{z}}{\mathbf{z}'V\mathbf{z}} + 2\ \frac{\mathbf{z}'\mathbf{z}}{w}\ \frac{r^{(1)}(\frac{\mathbf{z}'V\mathbf{z}}{w})}{r(\frac{\mathbf{z}'V\mathbf{z}}{w})}\right)\right].$$

Next, taking expected values in the order $E_{\mathbf{z},V}E_w$ and using Lemma 2 to integrate by parts over w yields

$$E\left[h^2\left(\frac{\mathbf{z}'V\mathbf{z}}{w}\right)\mathbf{z}'\mathbf{z}\right] = E\left[\frac{\mathbf{z}'\mathbf{z}}{(\mathbf{z}'V\mathbf{z})^2}\ w^2\ r^2\left(\frac{\mathbf{z}'V\mathbf{z}}{w}\right)\right]$$

$$(3.7) \qquad = 2E\left[\frac{\mathbf{z}'\mathbf{z}\ r^2\left(\frac{\mathbf{z}'V\mathbf{z}}{w}\right)}{(\mathbf{z}'V\mathbf{z})^2}\left\{\left(\frac{m-p+2}{2}\right)w - 2d\left(\frac{\mathbf{z}'V\mathbf{z}}{w}\right)\right\}\right]$$

where $d(u) = ur(u)r^{(1)}(u)$. Substituting (3.6) and (3.7) into (3.5), and using the assumption (2.2(ii)) that $r(u)$ is nondecreasing in $u, u \geq 0$, yields

$$(3.8) \qquad \omega(\tau) \geq \omega^*(\tau),$$

where

$$\omega^*(\tau) = 2E\left[h\left(\frac{\mathbf{z}'V\mathbf{z}}{w}\right)\left\{tr\ V^{-1} - k\ \frac{\mathbf{z}'\mathbf{z}}{\mathbf{z}'V\mathbf{z}}\right\}\right]$$

and

$$k = \left(\frac{m-p+2}{2}\right)\left(\sup_{u\geq 0} r(u)\right) + 2.$$

Note that (2.5) is equivalent to

$$(3.9) \qquad k < p.$$

Since w is independent of $\mathbf{z}$ and V, taking the expectation over w in $\omega^*(\tau)$, yields

$$\omega^*(\tau) = E\left[H(\mathbf{z}'V\mathbf{z})\left\{tr\ V^{-1} - k\ \frac{\mathbf{z}'\mathbf{z}}{\mathbf{z}'V\mathbf{z}}\right\}\right]$$

where

$$H(\mathbf{z}'V\mathbf{z}) = E_w\left[2h\left(\frac{\mathbf{z}'V\mathbf{z}}{w}\right)\right] = \frac{2}{\mathbf{z}'V\mathbf{z}}\,E_w\left[w\;r\left(\frac{\mathbf{z}'V\mathbf{z}}{w}\right)\right]$$

$$(3.10) \qquad \equiv \frac{2}{\mathbf{z}'V\mathbf{z}}R(\mathbf{z}'V\mathbf{z}).$$

Since $r(u)$ is nondecreasing in $u, u \geq 0$, it is easy to see that $R(u) = 2uH(u)$ is also nondecreasing in $u, u \geq 0$.

Simplification

Let

$$T = \begin{pmatrix} t_{11} & T_{12} \\ \mathbf{0} & T_{22} \end{pmatrix}$$

satisfy

$$V = T'T, \quad T_{22} \text{ symmetric .}$$

Also let

$$\mathbf{s} = (s_1, \mathbf{s}_2')' = T\mathbf{z}$$

where s_1 is a scalar. Since $TV^{-1}T' = I_p$,

$$(3.11) \qquad \mathbf{s}|T \sim \text{ MVN } (T\gamma, I_p) = \text{ MVN } ((\tau t_{11}, 0, \ldots, 0)', I_p).$$

Lemma 3.

$$(i)\; t_{11}^2 \sim \chi_m^2,\; T_{12}' \sim \text{ MVN } (\mathbf{0}, I_{p-1}),$$

$$(ii)\; T_{22}^2 \sim \mathcal{W}_{p-1}(m-1, I_{p-1}),$$

$$(iii)\; s_1|t_{11} \sim N(\tau^{1/2}t_{11}, 1)$$

$$(iv)\; \mathbf{s}_2 \sim \text{ MVN } (\mathbf{0}, I_{p-1})$$

and

$$(v)\; (s_1,\; t_{11}),\; \mathbf{s}_2,\; T_{12},\; T_{22} \text{ are mutually independent.}$$

Proof. The results (i) and (ii), and the independence of t_{11}, T_{12}, T_{22} follow directly from (3.4) and the definition of T (Muirhead, 1982). The results (iii), (iv) and (v) are then direct consequences of (3.11). □

Note that

$$T^{-1} = \begin{pmatrix} t_{11}^{-1} & -t_{11}^{-1} \quad T_{12} \quad T_{22}^{-1} \\ \mathbf{0} & T_{22}^{-1} \end{pmatrix}.$$

Further, it follows from Lemma 3 that

$$E[T_{12}] = \mathbf{0},\ E[T_{12}T_{12}'] = I_{p-1},\ E[T_{22}^{-2}] = \frac{1}{m-p-1}\, I_{p-1},$$

and that $(t_{11},\ \mathrm{s})$, T_{12} and T_{22} are mutually independent.

Thus,

$$\omega^*(\tau) = E_{(t_{11},\ \mathrm{s})} E_{T_{12},\ T_{22}} \left[H(\mathrm{s}'\mathrm{s}) \left\{ tr\left[T^{-1}(T')^{-1}\right] - k\ \frac{\mathrm{s}'(T')^{-1}T^{-1}\mathrm{s}}{\mathrm{s}'\mathrm{s}} \right\}\right]$$

$$(3.12) \quad = E_{(t_{11},\ \mathrm{s})} \left[H(\mathrm{s}'\mathrm{s}) \left\{ tr\left[E_{T_{12},\ T_{22}}(T^{-1}(T')^{-1})\right] - k\ \frac{\mathrm{s}'[E_{T_{12},\ T_{22}}((T')^{-1}T^{-1})]\mathrm{s}}{\mathrm{s}'\mathrm{s}} \right\}\right]$$

$$= E_{(t_{11},\ \mathrm{s})} \left[H(\mathrm{s}'\mathrm{s}) \left\{ \frac{(m-2)t_{11}^{-2} + (p-1)}{m-p-1} - k\left[\frac{s_1^2 t_{11}^{-2}}{\mathrm{s}'\mathrm{s}} + \frac{\mathrm{s}_2'\mathrm{s}_2(1+t_{11}^{-2})}{\mathrm{s}'\mathrm{s}(m-p-1)}\right]\right\}\right].$$

From (3.9) and (3.12), after some algebraic simplification, it follows that

$$(3.13) \qquad \omega^*(\tau) \geq \tilde{w}(\tau),$$

where

$$\tilde{w}(\tau) = E_{(t_{11},\ \mathrm{s})} \left[H(\mathrm{s}'\mathrm{s}) \left\{ \frac{(m-2)t_{11}^{-2} + (p-1)}{m-p-1} - p\left[\frac{s_1^2 t_{11}^{-2}}{\mathrm{s}'\mathrm{s}} + \frac{\mathrm{s}_2'\mathrm{s}_2(1+t_{11}^{-2})}{\mathrm{s}'\mathrm{s}(m-p-1)}\right]\right\}\right]$$

$$= E_{(t_{11},\ \mathrm{s})} \left[\frac{H(\mathrm{s}'\mathrm{s})}{m-p-1} \left(\frac{p s_1^2}{\mathrm{s}'\mathrm{s}} - 1 \right) \left(1 - (m-p-2)t_{11}^{-2}\right)\right]$$

It follows from Lemma 3 that

$$s_1^2 | t_{11} \sim \chi_1^2\left(\frac{t_{11}^2\tau}{2}\right) \sim \chi_{2J+1}^2$$

where $\chi_\nu^2(\delta^2)$ denotes a noncentral chi-squared random variable with noncentrality parameter δ^2 and ν degrees of freedom, and

$$J \sim \text{ Poisson } \left(\frac{t_{11}^2\tau}{2}\right).$$

(The representation of a noncentral chi-squared distribution as a Poisson mixture of central chi-squared distributions is a well-known result.) Let $f_\nu(\cdot)$ be the density function of a χ^2_ν random variable. It then follows from Lemma 3 that the joint density of s_1^2, $\mathbf{s}_2'\mathbf{s}_2$, and t_{11}^2 is given by

$$f(s_1^2,\ \mathbf{s}_2'\mathbf{s}_2,\ t_{11}^2) = \sum_{j=0}^{\infty} \left(\frac{t_{11}^2\tau}{2}\right)^j \left(\frac{e^{-\frac{1}{2}t_{11}^2\tau}}{j!}\right) f_{2j+1}(s_1^2) f_{p-1}(\mathbf{s}_2'\mathbf{s}_2) f_m(t_{11}^2). \tag{3.14}$$

The following is a well known distributional fact.

Lemma 4. If $x_i \sim \chi^2_{\nu_i}$, $i = 1,\ 2$, and $x_1,\ x_2$ are independent, then

$$\frac{x_1}{x_1+x_2} \text{ and } x_1 + x_2 \text{ are independent}$$

and

$$x_1 + x_2 \sim \chi^2_{\nu_1+\nu_2},\quad \frac{x_1}{x_1+x_2} \sim Beta\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right).$$

Let

$$h_j = E\left[H\left(\chi^2_{p+2j}\right)\right]. \tag{3.15}$$

Using (3.14) and (3.15) and Lemma 4, it is straightforward to show that

$$\begin{aligned}
\tilde{\omega}(\tau) &= \frac{p-1}{m-p-1}\sum_{j=0}^{\infty} h_j \left(\frac{2j}{2j+p}\right) E_{t_{11}}\left[\left(1-(m-p-2)\ t_{11}^{-2}\right)\left(\frac{t_{11}^2\tau}{2}\right)^j \frac{e^{-\frac{1}{2}t_{11}^2\tau}}{j!}\right] \\
&= \frac{p-1}{m-p-1}\sum_{j=0}^{\infty} q(j|(1+\tau)^{-1}\tau)h_j\left(\frac{2j}{2j+p}\right)\left(1-\frac{m-p-2}{m-2+2j}\ (1+\tau)\right) \\
&= \frac{(p-1)\tau}{m-p-1}\sum_{j=1}^{\infty} q(j-1|(1+\tau)^{-1}\tau)h_j\left(1-\frac{m+2j-2}{p+2j}\frac{\tau}{1+\tau}\right) \\
&= \frac{(p-1)\tau}{m-p-1}\sum_{i=0}^{\infty} q(i|(1+\tau)^{-1}\tau)h_{i+1}\left(1-\frac{m+2i}{p+2i+2}\frac{\tau}{1+\tau}\right).
\end{aligned}$$

where

$$q(i|\xi) = \left[\frac{\Gamma(\frac{m}{2}+i)}{\Gamma(\frac{m}{2})i!}\right] \xi^i (1-\xi)^{\frac{m}{2}}, \; i = 0, \; 1, \ldots.$$

Final Argument

Note that $q(i|\xi)$ is the mass function of the negative binomial distribution with parameters ξ and $m/2$. Hence, we can apply Hwang's discrete analog of integration by parts (Hwang, 1982, Equation (2.1)) to obtain

$$\begin{aligned} \tilde{\omega}(\tau) &= \frac{(p-1)\tau}{m-p-1} \sum_{i=0}^{\infty} q\left(i|\frac{\tau}{1+\tau}\right) [h_{i+1}] \left[1 - \frac{m+2i}{p+2i+2} \frac{\tau}{1+\tau}\right] \\ &= \frac{(p-1)\tau}{m-p-1} \sum_{i=0}^{\infty} q\left(i|\frac{\tau}{1+\tau}\right) \left[h_{i+1} - h_i \frac{2i}{p+2i}\right]. \end{aligned} \tag{3.16}$$

Lemma 5. For all $i = 0, 1, 2, \ldots,$

$$h_{i+1} \geq h_i \left[\frac{2i+p-2}{p+2i}\right] \geq h_i \left(\frac{2i}{p+2i}\right).$$

<u>Proof</u>. From (3.10) and (3.15),

$$h_{i+1} = E\left[H(\chi^2_{p+2i+2})\right] = E\left[\frac{R(\chi^2_2 + \chi^2_{p+2i})}{\chi^2_2 + \chi^2_{p+2i}}\right],$$

where χ^2_2 and χ^2_{p+2i} are independent chi-squared random variables with the degrees of freedom indicated in the notation. Since $R(u)$ is nondecreasing in $u \geq 0$,

$$\begin{aligned} &\frac{R(\chi^2_2 + \chi^2_{p+2i})}{\chi^2_2 + \chi^2_{p+2i}} - \frac{R(\chi^2_{p+2i})}{\chi^2_{p+2i}} \\ &= \frac{-\chi^2_2}{\chi^2_{p+2i}} \frac{R(\chi^2_2 + \chi^2_{p+2i})}{\chi^2_2 + \chi^2_{p+2i}} + \frac{\left[R(\chi^2_2 + \chi^2_{p+2i}) - R(\chi^2_{p+2i})\right]}{\chi^2_{p+2i}} \\ &\geq \frac{-\chi^2_2}{\chi^2_{p+2i}} \frac{R(\chi^2_2 + \chi^2_{p+2i})}{\chi^2_2 + \chi^2_{p+2i}}. \end{aligned}$$

Applying Lemma 4,

$$h_{i+1} - h_i = E\left[\frac{R(\chi_2^2 + \chi_{p+2i}^2)}{\chi_2^2 + \chi_{p+2i}^2}\right] - E\left[\frac{R(\chi_{p+2i}^2)}{\chi_{p+2i}^2}\right]$$

$$\geq -E\left[\frac{\chi_2^2}{\chi_{p+2i}^2}\left(\frac{R(\chi_2^2 + \chi_{p+2i}^2)}{\chi_2^2 + \chi_{p+2i}^2}\right)\right]$$

$$= -E\left[\frac{\chi_2^2}{\chi_{p+2i}^2}\right] E\left[\frac{R(\chi_2^2 + \chi_{p+2i}^2)}{\chi_2^2 + \chi_{p+2i}^2}\right]$$

$$= -\frac{2}{p+2i-2}\, h_{i+1}$$

or

$$h_{i+1} \geq h_i \frac{p+2i-2}{p+2i}, \quad i = 0, 1, 2, \ldots.$$

Since $h_i \geq 0$ all i and

$$\frac{p+2i-2}{p+2i} \geq \frac{2i}{p+2i}$$

when $p \geq 2$, the second inequality in the assertion of the lemma immediately follows. □

Lemma 5 and (3.16) shows that $\tilde{\omega}(\tau) \geq 0$. It now follows from (3.8) and (3.13) that $w(\tau) \geq 0$, and the proof of Theorem 1 is complete.

4. Improvement in Variance

Perhaps the simplest examples of the class of estimators $\mathbf{b}_h$ of β covered by Theorem 1 are those in which $r(u)$ is a constant, $r(u) = c$. In this case $r^{(1)}(u) = 0$ for all $u \geq 0$, so that the inequality (3.8) is actually an equality. As Stein had already noted (Stein, 1960, Section 6), among estimators of the form

$$\mathbf{b}_c = \left(1 - \frac{c}{\hat{\mathbf{b}}' W_{\mathbf{xx}} \hat{\mathbf{b}}}\right)\hat{\mathbf{b}},$$

the estimator

$$\mathbf{b}^* = \left(1 - \frac{(p-2)}{(n-p+1)(\hat{\mathbf{b}}' W_{\mathbf{xx}} \hat{\mathbf{b}})}\right)\hat{\mathbf{b}} \tag{4.1}$$

has minimum risk at $\beta = \mathbf{0}$. Recall that $\hat{\mathbf{b}}$ and all estimators $\mathbf{b}_h$ of the form (2.1) have risk depending on the parameters μ, Σ only through $\tau = (1 - \rho^2_{y\cdot\mathbf{x}})^{-1}\, \rho^2_{y\cdot\mathbf{x}}$ and that $\tau = 0$ if and only if $\beta = \mathbf{0}$. The risk of $\mathbf{b}^*$ when $\beta = \mathbf{0}$ ($\tau = 0$) is found by Stein to be

$$R(\mathbf{b}^*; 0) = \frac{2(n-1)}{(n-p-2)(n-p+1)}. \tag{4.2}$$

The risk of $\hat{\mathbf{b}}$ is constant for all τ;

$$R(\hat{\mathbf{b}}; \tau) = \frac{p}{n-p-2}. \tag{4.3}$$

The relative saving in variance of $\overline{y}(\mathbf{b}^*)$ versus $\overline{y}(\hat{\mathbf{b}})$ is (see (1.7))

$$RS(\tau) = \frac{\mathrm{var}[\overline{y}(\hat{\mathbf{b}})] - \mathrm{var}[\overline{y}(\mathbf{b}^*)]}{\mathrm{var}\ \overline{y}(\hat{\mathbf{b}})} = \frac{R(\hat{\mathbf{b}};\tau) - R(\mathbf{b}^*;\tau)}{1 + R(\hat{\mathbf{b}};\tau)}. \tag{4.4}$$

Because of the term n^{-1} in the variances of $\overline{y}(\hat{\mathbf{b}})$ and $\overline{y}(\mathbf{b}^*)$, their variances are typically small. Gains in accuracy of $\overline{y}(\mathbf{b}^*)$ over $\overline{y}(\hat{\mathbf{b}})$ as an estimator of μ_y are more easily seen by calculating relative variances.

From (4.2), (4.3), (4.4), when $\tau = 0$ (i.e. $\beta = \mathbf{0}$)

$$RS(0) = \frac{p - \frac{2(n-1)}{n-p-1}}{n-2}.$$

The relative saving will be large when p is large relative to n. For example, when $n = 20$ and $p = 10$, $RS(0) = 0.32$ or 32%. As noted by Bauer (1987), there are usually a large number of possible control variables available in a simulation study. Because adding extra control variables may increase the multiple correlation $\rho_{y\cdot\mathbf{x}}$, there is the temptation to use as many of these variables as possible. Since the number n of repetitions of the simulation experiment is often fixed by cost considerations, it is not unusual for the number p of control variables to be large relative to n. [However, remember that Theorem 1 requires that $n > p + 2$.]

The relative savings in variance for $\overline{y}(\mathbf{b}^*)$ versus $\overline{y}(\hat{\mathbf{b}})$ is greatest for τ near 0, and least when τ is large ($\rho^2_{y\cdot\mathbf{x}}$ is near 1). One can use the methods of Section 3 to find a series representation for $R(\mathbf{b}^*;\tau)$. However, the present discussion is only intended to exhibit the virtues of the class of estimators $\overline{y}(\mathbf{b}_h)$, not to give the properties of any particular estimator in this class.

Indeed, $\mathbf{b}^*$ is inadmissible as an estimator of β, and consequently $\overline{y}(\mathbf{b}^*)$ is inadmissible as an estimator of μ_y. Standard arguments show that $\mathbf{b}_h$ with

$$h(u) = \begin{cases} 0, & \text{if } u \le \frac{p-2}{n-p+1}, \\ \frac{p-2}{(n-p+1)u}, & \text{if } u > \frac{p-2}{n-p+1}, \end{cases}$$

dominates $\mathbf{b}^*$ in risk. (This is the familiar James-Stein modification.) This last estimator is also known to be inadmissible, but serves as an adequate approximation to an admissible estimator. Unfortunately its risk function, even at $\tau = 0$, is difficult to evaluate.

The present author believes that a Bayes, or robust Bayes (Berger, 1984a), approach offers the most satisfactory approach to choosing an estimator $\mathbf{b}$ of β. If one wants an estimator which is equivariant under the transformations used at the start of Section 3, the class of estimators (2.1) provides a variety of reasonable choices, even when this class is restricted by the conditions (2.2) and (2.5). Although this paper has concentrated on the use of estimators of β in simulation studies, Theorem 1 also has uses in prediction of future y-values (Stein, 1960; Copas, 1983), and in cases where point estimation of β is the final goal of inference.

In simulation studies using control variates, investigators typically report a confidence interval for μ_y, rather than merely giving a point estimator. It is likely (see Hwang and Casella, 1982) that replacing $\overline{y}(\hat{\mathbf{b}})$ in such intervals by any of the estimators $\overline{y}(\mathbf{b}_h)$ covered by Theorem 1 will increase the coverage probability of such intervals, particularly when $\rho^2_{y\cdot\mathbf{x}}$ is small. However, the minimal coverage probability will still be that stated for the interval centered at $\overline{y}(\hat{\mathbf{b}})$, so that any gains in coverage probability will go unreported. Consequently, a preferable approach would be to combine construction of such intervals with a report of estimated coverage probability using methods similar to those of Lu (1987). Similar comments pertain to the construction of confidence regions for β, confidence intervals for linear combinations of β, and prediction intervals for future values of y. Research on such methodology is currently in progress.

Finally, it should be mentioned that the problem treated in this paper is a special case of more general problems in which mean vectors, matrices of means, or slope vectors or matrices are to be estimated, and where prior knowledge is available concerning the value(s) of certain elements (or linear combinations of elements) of these vectors or matrices. Examples of such problems include GMANOVA (Gleser and Olkin, 1966, 1970) and seemingly unrelated regression (SUR) problems (Zellner, 1962). See also Kariya (1985) for an overview of these problems which reveals their basic similarities and structure. Such problems offer the potential for combining the theory of the present paper with the more familiar shrinkage methodology of Stein (1981), Efron and Morris (1972) and other authors.

REFERENCES

Bauer, K. W., Jr. (1987). *Control Variate Selection for Multiresponse Simulation.* Ph.D. Dissertation, Department of Industrial Engineering, Purdue University.

Berger, J. (1984). The robust Bayesian viewpoint. In *Robustness in Bayesian Statistics* (J. Kadane, ed.). North-Holland:Amsterdam.

Copas, J. B. (1983). Regression, prediction and shrinkage. *J. R. Statist. Soc. B* **45**, 311–354.

Efron, B. and Morris, C. (1972). Empirical Bayes on vector observations - An extension of Stein's method. *Biometrika* **59**, 335–347.

Gleser, L. J. (1987). Improved estimators of mean response in simulation when control variates are used. Technical Report, Mimeo Series #87-32, Department of Statistics, Purdue University.

Gleser, L. J. and Olkin, I. (1966). A k-sample regression model with covariance. In *Multivariate Analysis* (P. R. Krishnaiah, ed.). Academic Press:New York.

Gleser, L. J. and Olkin, I. (1970). Linear models in multivariate analysis. In *Essays in Probability and Statistics* (R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, K. J. C. Smith, eds.). University of North Carolina Press:Chapel Hill, N.C.

Hwang, J. T. (1982). Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases. *Ann. Statist.* **10**, 857–867.

Hwang, J. T. and Casella, G. (1982). Minimax confidence sets for the mean vector of a multivariate normal distribution. *Ann. Statist.* **10**, 868–881.

Kariya, T. (1985). *Testing in the Multivariate General Linear Model.* Kinokuniya Co. Ltd.:Tokyo.

Kleijen, J. P. C. (1975). *Statistical Techniques in Simulation, Part 1 and 2.* Marcel Dekker:New York.

Lu, Kun-Liang (1987). *The Estimated Loss Frequentist Approach.* Ph.D. Dissertation, Department of Statistics, Purdue University.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* John Wiley:New York.

Sclove, S. L., Morris, C. and Radhakrishnan, R. (1972). Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.* **43**, 1481–1490.

Stein, C. (1960). Multiple regression. In *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling.* (Olkin, I.; Ghurye, S.G.; Hoeffding, W.; Madow, W. G.; Mann, H. B., eds). Stanford University Press:Stanford, California.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135–1151.

Wilson, J. R. (1984). Variance reduction techniques for digital simulation. *Amer. J. of Math. and Man. Sci.* **1**, 227–312.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regression and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57**, 348–368.