COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station

New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 786

BEST MEDIAN UNBIASED ESTIMATION IN LINEAR REGRESSION WITH BOUNDED ASYMMETRIC LOSS FUNCTIONS

Donald W. K. Andrews and Peter C. B. Phillips*

March 1986

Author's Footnote

*Donald W. K. Andrews is Assistant Professor, Cowles Foundation for Research in Economics, Department of Economics, Yale University, P.O. Box 2125, Yale Station, New Haven, CT 06520; and Peter C. B. Phillips is Professor, Cowles Foundation for Research in Economics, Department of Economics, Yale University, P.O. Box 2125, Yale Station, New Haven, CT 06520. The first author received support from the Sloan Foundation, through a Research Fellowship, and from the National Science Foundation, via grant number SES-8419789. The authors would like to thank two referees for helpful comments and reference to the paper by Hwang (1985).

Key Words

Generalized least squares, elliptically symmetric distribution, restricted parameter space, minimum risk, variance estimation.

EXTENDED ABSTRACT

This paper considers optimal median unbiased estimation in a linear regression model with the distribution of the errors lying in a sub-class of the elliptically symmetric distributions. The generalized least squares (GLS) estimator is shown to be best for any monotone loss function, i.e., any loss function that is nondecreasing as the magnitude of under-estimation or over-estimation increases. This includes bounded asymmetric loss functions. For the same loss functions, a restricted GLS estimator is shown to be best when the estimand is known to lie in an interval. For the case of normal errors, a best median unbiased estimator of the error variance σ^2 is given, for the cases of restricted and unrestricted parameter spaces. This estimator differs from the sample variance s^2. In comparison with best mean unbiased estimators of regression and variance parameters, the best median unbiased estimators considered here take advantage of restrictions on the parameter space, and are optimal with respect to a much wider class of loss functions--in particular, both bounded and unbounded loss functions.

The choice of median unbiasedness, as opposed to mean unbiasedness, is not crucial when deriving an optimality result for the estimation of regression parameters when the model has elliptically symmetric errors, provided the parameter space is unrestricted, or is restricted only by linear constraints. The reason is that many estimators considered in the literature have symmetric distributions about the estimand in this context, and hence, are both median and mean unbiased if their expectation exists. (Proper Bayes and shrinkage estimators are the two main classes of estimators that do not have symmetric distributions and are neither mean nor median unbiased.)

On the other hand, if the parameter space of the regression parameters is restricted by nonlinear constraints on the parameters, then the mean unbiasedness condition becomes much more restrictive than median unbiasedness. This occurs because estimators that take advantage of the restrictions on the parameters generally are (mean) biased. Median unbiased estimators, however, can be adjusted to take account of restrictions without losing their property of median unbiasedness. Thus, our use of the condition of median unbiasedness, rather than mean unbiasedness, is of little consequence when the parameter space is unrestricted, and is a distinct advantage when the parameter space is restricted by nonlinear constraints on the parameters.

The class of error distributions that we consider consists of distributions that are consistent with elliptical symmetry for any sample size. Such distributions are rotated variance mixtures of multivariate normal distributions (and hence, include multivariate normal distributions). An example of a situation in which a non-normal elliptically symmetric error distribution may arise is the following. Consider the classical regression model based on an agricultural experiment. Suppose the dependent variable is crop yield, and the independent variables include fertilizer treatment. The error may be comprised of several factors including differential land quality. The seed for each plot is taken from the same stock. The quality of this stock may be viewed as the outcome of a random draw (with different points in time or different geographic origins of the stock yielding different draws). Conditional on the stock of seed used, it may be reasonable to assume that the errors have a normal distribution. Different stocks of seed may interact differently with the environment to yield different conditional variances of the errors. To make inferences that are valid for the population of seed stocks, then, one needs to treat the errors as

a variance mixture of normal distributions.

As a second example, consider a regression model with economic variables where the observations correspond to different firms in an industry observed at the same point in time. Suppose the errors are identically distributed across firms, and the state of the macro economy affects the size of the error variance for each firm. It may be reasonable to assume that the errors have normal distribution conditional on the state of the macro economy. If so, then one needs to treat the errors as a variance mixture of normals if one wishes to make inferences that are valid for different points in the business cycle.

These examples suggest that there are a number of situations in which it may be reasonable to assume that the errors have non-normal elliptically symmetric distributions. Of course, there are many additional situations for which the assumption of normality is appropriate.

The contents of this paper are organized as follows. Section 1 briefly reviews recent results by Kariya (1985) and Hwang (1985) that are related to the results given here. Section 2 shows that the GLS estimator is the best median unbiased estimator of the regression parameters for quite general loss functions, when the parameter space is unrestricted. Of note is the fact that this result holds without moment restrictions. Thus, the errors may have multivariate Cauchy distribution. Section 3 shows that a restricted GLS estimator is best median unbiased for a linear combination of the regression parameters, when that linear combination is restricted to lie in an interval. Certain other linear combinations of the parameter vector may be subject to arbitrary additional restrictions. Section 4 presents best median unbiased estimators of the error variance σ^2, as well as monotone functions of σ^2, when the errors are normally distributed.

If σ^2 is constrained to lie in a finite interval, the best estimator is a censored version of its unconstrained counterpart. When σ^2 is constrained only to be positive, the best median unbiased estimator is always larger than the best mean unbiased estimator s^2, and is approximately equal to s^2 calculated with its degrees of freedom reduced by .66. The final Section 5 gives proofs of the results. These make use of results due to Lehmann (1959) and Pfanzagl (1979).

1. KARIYA'S AND HWANG'S OPTIMALITY RESULTS FOR GLS

The Gauss-Markov Theorem states that for the linear regression model,

$$y = X\beta_0 + u_N \ , \quad E(u_N) = 0 \ , \quad \text{and} \quad \text{Cov}(u_N) = \sigma^2\Sigma \ , \tag{1}$$

the generalized least squares (GLS) estimator,

$$\hat{\hat{\beta}} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \ , \tag{2}$$

is the best linear unbiased estimator, in the sense that $c'\hat{\hat{\beta}}$ minimizes the mean squared error for estimation of $c'\beta_0$ for all fixed K-vectors c, provided Σ is known. Here X is an $N \times K$ fixed matrix of rank K, Σ is a positive definite $N \times N$ matrix, and $\beta_0 \in R^K$.

Two extensions of this result have appeared recently in the literature, see Kariya (1985) and Hwang (1985, Corollary 3.2). In this section, we briefly review these extensions.

A common criticism of the Gauss-Markov Theorem is that it only considers linear estimators. This has little or no justification. In contrast, Kariya's (1985) recent version of the Gauss-Markov Theorem allows for nonlinear estimators. The class of estimators he considers is

$$\mathcal{C}_1 = \{\hat{\beta} \mid \hat{\beta}(y) = C(e)y, \ C(e) \text{ is a } K \times N \text{ matrix-valued measurable function of } e \text{ such that } C(e)X = I_K \text{ for all } e \text{ and } E\|\hat{\beta}\|^2 \text{ exists}\} \ ,$$

where e is the N-vector of ordinary least squares (OLS) residuals, i.e., $e = y - Xb$, where $b = (X'X)^{-1}X'y$ is the OLS estimator. This class includes the class of linear unbiased estimators. It contains both nonlinear and biased estimators. A typical example of an estimator in $\mathcal{C}_1$ is the nonlinear feasible GLS estimator defined by $C(e) = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}$, where

the estimated covariance matrix $\hat{\Sigma}$ depends only on the OLS residuals e.

Kariya's optimality result for the class $\mathcal{C}_1$ is possible, because he considers a smaller class of error distributions than is considered in the Gauss-Markov Theorem. Let $\mathcal{F}_N^2$ be the class of distributions satisfying (1) such that when u_N is transformed into $\tilde{u}_N = \Sigma^{-1/2}u_N$, the distribution of $\tilde{u}_N$ is orthogonally invariant; that is, $\mathcal{L}(\Gamma\tilde{u}_N) = \mathcal{L}(\tilde{u}_N)$, where Γ is any $N \times N$ orthogonal matrix and $\mathcal{L}(\cdot)$ denotes the distribution of $\cdot$. $\mathcal{F}_N^2$ is the class of *elliptically symmetric* N-variate distributions with two moments finite (see Kelker (1970) and King (1980)). It contains the N-variate normal distribution, and N-variate exponential and t-distributions with three or more degrees of freedom (see Lord (1954), Dunnett and Sobel (1955), and Bennett (1961)).

Kariya shows that if $\mathcal{L}(u_N) \in \mathcal{F}_N^2$, then the GLS estimator is best in the class $\mathcal{C}_1$ in the sense of mean squared error. That is, for any $\hat{\beta} \in \mathcal{C}_1$ and all $c \in R^K$,

$$E(c'\hat{\beta} - c'\beta_0)^2 \geq E(c'\hat{\tilde{\beta}} - c'\beta_0)^2 = \sigma^2 c'(X'\Sigma^{-1}X)^{-1}c \ . \qquad (3)$$

As Kariya points out, this result has relevance whether or not Σ is known: If Σ is known it yields a best estimator; if Σ is unknown, it yields a sharp lower bound for the mean squared error of estimators in $\mathcal{C}_1$.

Although Kariya's class $\mathcal{C}_1$ is more general than the class of linear unbiased estimators, it is still quite restrictive. It excludes a wide variety of estimators in the literature that are unbiased in the present context. Such estimators include maximum likelihood for unknown Σ, robust M-, L-, R-, minimum distance, spectral, and adaptive estimators (see Andrews (1986) and the discussion in Section 2 below). In addition, Kariya's result only establishes optimality with respect to the squared error loss function.

Hwang (1985, Corollary 3.2) extends the Gauss-Markov Theorem in a different direction from that of Kariya. He generalizes the criterion of optimality considerably from mean squared error to risk under arbitrary symmetric monotone loss functions (defined below). The squared error loss function is of very special form, and exhibits the general qualitative features of unboundedness and symmetry. In many circumstances, this loss function is not very appropriate. Hence, it is important to see if the optimality of the GLS estimator is sensitive to this particular choice of loss function.

When the errors have multivariate normal distribution, it is known that the GLS estimator $c'\hat{\beta}$ is the best unbiased estimator of $c'\beta_0$ for any convex loss function (see Lehmann (1983, Theorem 3.4.3, p. 189)). This is a generalization of the standard UMVU result. It is quite useful, because it allows for asymmetric loss functions of fairly flexible shape, and does not impose linearity of the estimators. Unfortunately, the convexity condition implies unboundedness of the loss function, which may be inappropriate in many circumstances.

Hwang's result, on the other hand, imposes symmetry of the loss functions, but otherwise allows for quite general shape, including boundedness. For estimation of $c'\beta_0$, he considers non-negative loss functions $L(c'\hat{\beta} - c'\beta_0)$ that are symmetric about zero and nondecreasing in $|c'\hat{\beta} - c'\beta_0|$. (In fact, Hwang's (1985) results carry through unchanged for loss functions $L(\beta_0, c'\hat{\beta} - c'\beta_0)$ that are symmetric about zero in their second argument and nondecreasing in $|c'\hat{\beta} - c'\beta_0|$, for each value of their first argument $\beta_0 \in R^K$. This extension can be important, because the magnitude of the loss attributable to over- or under-estamation by a fixed amount often depends on the true value of the parameter.)

Hwang (1985, Theorem 2.3) shows that his class of loss functions is sufficiently general that given two estimators $c'\hat{\beta}_1$ and $c'\hat{\beta}_2$, the risk of $c'\hat{\beta}_1$ is less than or equal to that of $c'\hat{\beta}_2$ for all symmetric monotone loss functions if and only if

$$|c'\hat{\beta}_1 - c'\beta_0| \overset{ST}{\leq} |c'\hat{\beta}_2 - c'\beta_0| , \tag{4}$$

where $\overset{ST}{\leq}$ denotes "stochastically less than or equal to." Thus, optimality under Hwang's class of loss functions is a strong result. The only clear deficiency is the restriction to *symmetric* loss functions.

To show the optimality of GLS under symmetric monotone loss functions, Hwang assumes the errors have elliptically symmetric distributions and maintains the restriction to linear estimators that is used in the Gauss-Markov Theorem. In addition, he assumes that the estimators are either median unbiased or (mean) unbiased. In the former case, the error distributions are not subject to any moment restrictions, i.e., $\mathcal{L}(u_N) \in \mathcal{F}_N^0$. (The superscript 0 denotes the assumed number of well-defined moments. For distributions in $\mathcal{F}_N^0$ that have infinite variances, Σ does not satisfy (1), since no covariance matrix exists. In this case, Σ is just the characteristic matrix that achieves spherical symmetry in the transformed coordinates.) In the latter case, the error distributions are assumed to have one moment well-defined, i.e., $\mathcal{L}(u_N) \in \mathcal{F}_N^1$. Thus, Hwang's error assumptions are stronger than those of the Gauss-Markov theorem with respect to the range of distributions with finite variances, but are more general in terms of moment restrictions.

Under these assumptions, Hwang shows that the GLS estimator $c'\hat{\hat{\beta}}$ of $c'\beta_0$ is best in the class of linear unbiased (or median unbiased) estimators for all symmetric monotone loss functions. Thus,

$$|c'\hat{\beta} - c'\beta_0| \overset{ST}{\leq} |c'\hat{\beta} - c'\beta_0| \ , \tag{5}$$

for all linear unbiased estimators $c'\hat{\beta}$.

This is an interesting result, but it suffers greatly from the arbitrary restriction to linear estimators. Also, the restriction to symmetric loss functions may be objectionable.

2. A STRONG OPTIMALITY RESULT FOR GLS

Each of the optimality results discussed above is less general than desirable due to the class of loss functions considered and/or the class of estimators considered. For example, none of these results allows for bounded asymmetric loss functions. Further, the results of Kariya and Hwang arbitrarily restrict the class of estimators beyond the restriction due to unbiasedness or median unbiasedness (which itself may be subject to criticism). The result we present here removes these restrictions.

For estimation of $c'\beta_0$, given $c \in R^K$, we consider loss functions $L(\beta_0, c'\hat{\beta} - c'\beta_0)$ that are subject only to the condition that loss is non-decreasing in $c'\hat{\beta} - c'\beta_0$ for $c'\hat{\beta} - c'\beta_0 > 0$, and non-increasing in $c'\hat{\beta} - c'\beta_0$ for $c'\hat{\beta} - c'\beta_0 < 0$, for each value of its first argument $\beta_0 \in R^K$. Such loss functions are called *monotone*. They are considered by Lehmann (1959, p. 83) and Pfanzagl (1979).

For particular choices of monotone loss functions the risk of an estimator $c'\hat{\beta}$ is given by $P(-d_1 \leq c'\hat{\beta} - c'\beta_0 \leq d_2)$, for any $d_1, d_2 \geq 0$. Thus, an estimator that is optimal with respect to the class of monotone loss functions has distribution more concentrated about the estimand than any other estimator considered. This is a strong optimality result.

The argument of Hwang (1985, Theorem 2.3) can be used to show that for two estimators $c'\hat{\beta}_1$ and $c'\hat{\beta}_2$, the risk of $c'\hat{\beta}_1$ is less than or

equal to that of $c'\hat{\beta}_2$ for all monotone loss functions if and only if

$$(c'\hat{\beta}_1 - c'\beta_0)_+ \overset{ST}{\leq} (c'\hat{\beta}_2 - c'\beta_0)_+ \quad \text{and} \quad (c'\hat{\beta}_1 - c'\beta_0)_- \overset{ST}{\leq} (c'\hat{\beta}_2 - c'\beta_0)_- \ , \quad (6)$$

where $(\cdot)_+$ and $(\cdot)_-$ denote the positive and negative parts of $\cdot$, i.e., for $\lambda \in R$, $(\lambda)_+ = \max\{\lambda,0\}$ and $(\lambda)_- = \max\{-\lambda,0\}$. If an estimator $\hat{\beta}_1$ satisfies (6) for all $\hat{\beta}_2$ in a designated class, we say that $\hat{\beta}_1$ is *stochastically best* in this class of estimators. This is a stronger result than optimality with respect to Hwang's stochastic condition (4).

A particular monotone loss function that may be of interest is the function $L(\beta_0, s) = s^2/(1+\lambda s^2)$, for $\lambda > 0$, where $s = c'\hat{\beta} - c'\beta_0$. This loss function is bounded, yet for small λ it is close to the common squared error loss function except when s^2 is large. Of course, the (unbounded) squared error loss function is also a monotone loss function.

The class of error distributions that we consider is slightly less general than the class $\mathcal{F}_N^0$ of elliptically symmetric N-variate distributions centered at the origin. In most applications of the linear regression model, the properties of the errors are not specific to the sample size under consideration. In particular, if an assumption such as elliptical symmetry of the errors is reasonable for sample size n equal to some N, then it is necessarily reasonable for sample size n equal to $N-1, N-2, \ldots, 1$, and usually also is reasonable for sample sizes $N+1, N+2, \ldots$. This being the case, it is not unduly restrictive to consider the sub-class of error distributions of $\mathcal{F}_N^0$ given by

$$\mathcal{C}_N = \{\mathcal{L}(u_N) \in \mathcal{F}_N^0 : \mathcal{L}(u_n) \in \mathcal{F}_n^0 \text{ for } n = 1, 2, \ldots\} \ ,$$

where u_n denotes the vector of errors $(u_{(1)}, u_{(2)}, \ldots, u_{(n)})'$ when the

sample size is n. That is, $\mathcal{C}_N$ contains all distributions of the first N errors that can be generated by errors $(u_{(1)}, u_{(2)}, \ldots, u_{(n)})'$ that have elliptically symmetric distributions for any sample size $n = 1, 2, \ldots$.

Distributions in $\mathcal{C}_N$ are called *consistent elliptically symmetric* (CES) N-variate distributions, where the adjective "consistent" refers to the fact that the distributions are consistent with elliptical symmetry for any sample size n. Since $\mathcal{C}_N$ is not restricted by moment conditions, it contains distributions with infinite variances and undefined means. In particular, $\mathcal{C}_N$ contains the N-variate normal, exponential, and t-distributions, including the N-variate Cauchy distribution.

By Theorem 10 of Kelker (1970), $\mathcal{L}(u_N) \in \mathcal{C}_N$ if and only if the distribution of $\Sigma^{-1/2}u_N$ is a variance mixture of N independent identically distributed mean zero normal random variables (with non-negative mixing density). Thus, CES distributions can be constructed and characterized quite simply.

In comparison with the error distributions considered in the Gauss-Markov Theorem, the class of CES distributions restricts the range of distributions with finite variances considerably. On the other hand, this restriction weakens the conditions of unbiasedness and median unbiasedness substantially, as we now shall see.

The class of estimators that we consider consists of all *median unbiased* estimators. By definition, an estimator $c'\hat{\beta}$ of $c'\beta_0$ is median unbiased if

$$P(c'\hat{\beta} \geq c'\beta_0) \geq 1/2 \quad \text{and} \quad P(c'\hat{\beta} \leq c'\beta_0) \geq 1/2 . \tag{7}$$

If $P(c'\hat{\beta} = c'\beta_0) = 0$, as is usually the case, then this condition simplifies to $P(c'\hat{\beta} > c'\beta_0) = P(c'\hat{\beta} < c'\beta_0) = 1/2$.

In the present context, the class of median unbiased estimators is very large--much larger than the class of unbiased or median unbiased estimators in the Gauss-Markov set up. The reason is that u_N is symmetrically distributed about the zero vector (i.e., $\mathcal{L}(u_N) = \mathcal{L}(-u_N)$) when it has an elliptically symmetric distribution. Thus, all estimators $\hat{\beta}$ that are odd functions of the errors have distributions symmetric about β_0, and yield median unbiased estimators $c'\hat{\beta}$ of $c'\beta_0$, for all $c \in R^K$. As shown in Andrews (1986), this result applies to the majority of non-Bayesian, non-shrinkage estimators considered in the literature. It holds for a wide class of nonlinear estimators that are defined as solutions to maximization problems or systems of equations, where initial estimators may be employed. This includes iterated estimators. In particular, the following estimators are covered: feasible GLS, quasi-maximum likelihood, Huber M-, bounded-influence M-, L-, R-, minimum distance, spectral, band spectral, GEM (see Andrews (1983)), adaptive, one-step asymptotically efficient, and instrumental variable. Note that these estimators also are mean unbiased provided their expectation exists.

Our main result is the following theorem. Its proof makes use of a result of Lehmann (1959, pp. 80-83) for best median unbiased estimation in monotone likelihood ratio families of distributions that are indexed by a scalar parameter. A different proof of our result can be obtained by applying an extension of Lehmann's result due to Pfanzagl (1979).

The term "unique" is used in the Theorem to mean unique almost everywhere with respect to Lebesgue measure.

Theorem 1. *Consider the model* $y = X\beta_0 + u_N$, *where* $\beta_0 \in R^K$, X *is full rank, and* $\mathcal{L}(u_N) \in \mathcal{G}_N$. *(a) The* GLS *estimator* $c'\hat{\beta}$ *is the unique best median unbiased estimator of* $c'\beta_0$ *for any given* $c \in R^K$ *in the sense of uniformly minimum risk for any monotone loss function. (b) Equivalently, the* GLS *estimator* $c'\hat{\beta}$ *is the unique stochastically best median unbiased estimator of* $c'\beta_0$, *for any given* $c \in R^K$.

The proof is given in Section 5 below.

Comments: 1. The Theorem also holds if we restrict attention to errors with multivariate normal distributions. The requirement of median unbiasedness under the larger class of CES distributions is not driving the optimality result by eliminating estimators from consideration.

2. The GLS estimator has infinite risk for some loss functions and some error distributions in $\mathcal{G}_N$. The Theorem still has import in these circumstances, however, because it implies that every other median unbiased estimator also has infinite risk.

3. In some cases, the ultimate object of interest is not $c'\beta_0$ but a nonlinear function of $c'\beta_0$, say $h(c'\beta_0)$, because it has a particular interpretation or meaning in an underlying theoretical model. For example, we may want to estimate the logarithm of a regression parameter. If $h(\cdot)$ is a monotone function, then given Theorem 1, it is not hard to see that not only is $h(c'\hat{\beta})$ median unbiased, but it is the best median unbiased estimator for any monotone loss function (under the assumptions of the Theorem). This is a very convenient result, especially in light of the difficulties in obtaining best mean unbiased estimators of nonlinear functions of $c'\beta_0$. Such estimators do not equal $h(c'\hat{\beta})$, in general, and may not even exist.

4. For bounded loss functions, the risk of the GLS estimator is finite even when the errors have undefined means or infinite variances, e.g., as in the N-variate Cauchy case. Thus, we get the interesting result that situations exist in which the least squares estimator is strictly preferred over a wide variety of robust procedures, even though the errors may have no moments finite. This result is possible, because the errors are not independent, even if $\Sigma = I_N$, unless u_N has normal distribution. The optimality result depends heavily on the elliptically symmetric form of the underlying error distribution, as comparisons with results in the robustness literature clearly atest.

5. The class of estimators considered in Theorem 1 is much more general with respect to nonlinearity than is Kariya's (1985) class $\mathcal{C}_1$. It does not contain $\mathcal{C}_1$, however, because $\mathcal{C}_1$ includes some biased estimators. On the other hand, if the function $C(e)$ that defines Kariya's estimators is an even function of the OLS residuals e, then $\hat{\beta} = C(e)y$ is median unbiased for $\mathcal{L}(u_N) \in \mathcal{F}_N^0$, since it is an odd function of the errors. Given the assumed symmetry of u_N about the zero vector, the evenness of $C(e)$ arises quite naturally, and most (or all) estimators in $\mathcal{C}_1$ that have been considered in the literature satisfy this property.

Nevertheless, for an optimality result it is desirable to avoid any restriction on the class of estimators, if possible. If one wishes to include the biased estimators of Kariya in an optimality result, one can proceed as follows: Consider the estimators of Theorem 1 of Andrews (1986) where the assumption A1 is relaxed by requiring the function r to be even in only its first argument rather than its first three arguments. Call the collection of such estimators $\mathcal{C}_2$. The class $\mathcal{C}_2$ contains $\mathcal{C}_1$. One can show for $\beta_0 \in R^K$ and $\mathcal{L}(u_N) \in \mathcal{E}_N$, the GLS estimator $\hat{\hat{\beta}}$ is the best

estimator of β_0 in the class $\mathcal{C}_2$ in the sense of uniformly minimum risk for any symmetric convex loss function (see Andrews and Phillips (1985)). This result generalizes Kariya's, because it considers much wider classes of loss functions and estimators (although it imposes slightly different error assumptions).

6. The result of the Theorem can be extended to allow homogeneous or non-homogeneous linear restrictions on β_0, and to allow less than full rank X matrix (provided identifying linear side conditions on β_0 are specified). If β_0 is subject to inequality constraints, however, then Theorem 1 no longer holds, but a restricted GLS estimator can be shown to possess similar strong optimality properties, as the next section illustrates.

7. As stated, Theorem 1 does not cover the standard multivariate regression model. It is not difficult, however, to use the proof of Theorem 1 to establish an analogous result for this model. Such a result is important because the multivariate regression model is of considerable interest in econometrics, due to its application to demand systems, among others.

The multivariate regression model consists of T observations on g equations, and can be written as

$$\underset{(T\times g)}{Y} = \underset{(T\times m)}{Z}\ \underset{(m\times g)}{A_0} + \underset{(T\times g)}{U} ,$$

where Y, Z, A_0, and U are matrices of dependent variables, regressors, unknown parameters, and errors, respectively. The parameter matrix A_0 may contain zeroes and redundant elements, and hence, is assumed to satisfy $\text{vec}(A_0) = S\beta_0$, where S is a $gm \times p$ known selection matrix (with $p \leq gm$), β_0 is the vector of basic unknown parameters, $\beta_0 \in R^p$,

and $\text{vec}(\cdot)$ denotes the row by row vectorization operator. Equivalently, this model can be written as $y = X\beta_0 + u$, where $y = \text{vec}(Y)$, $X = (I_g \otimes Z)S$, and $u = \text{vec}(U)$. Write $U = (u_1, \ldots, u_T)'$. Suppose the error vectors $u_1, \ldots, u_T$ are independent across observations, and each error vector u_j has some elliptically symmetric distribution with $g \times g$ full rank characteristic matrix Ω_j and no probability mass at the origin. (The vectors $u_1, \ldots, u_T$ need not be identically distributed.) Let $\hat{\beta}$ denote the GLS estimator of β_0 given by equation (2) with $\Sigma = \text{diag}(\Omega_1, \ldots, \Omega_T)$. The above class of distributions of u does not equal $\mathcal{E}_{gT}$, and hence, Theorem 1 does not apply. Nevertheless, it is straightforward to alter the proof of Theorem 1 to show that the optimality results (a) and (b) of Theorem 1 hold for the GLS estimator $\hat{\beta}$ in this multivariate regression model.

3. OPTIMAL ESTIMATION WITH A RESTRICTED PARAMETER SPACE

In this section we discuss optimal estimation of $c'\beta_0$ when β_0 is subject to certain nonlinear restrictions. In particular, we consider the case where $c'\beta_0$ is known to lie in a (possibly infinite) interval that does not depend on β_0, and certain linear combinations of β_0, denoted $c_2'\beta_0, \ldots, c_K'\beta_0$, are restricted in any fashion not involving $c'\beta_0$. A simple example is when we wish to estimate some element of β_0 subject to the sole constraint that this element is positive or lies in $[0,1]$.

Suppose the only restriction on β_0 is that $c'\beta_0$ lies in a nondegenerate interval strictly contained in R. The best estimator of $c'\beta_0$ from a sub-class of mean unbiased estimators is the GLS estimator that ignores the constraints, according to the Gauss-Markov Theorem, Kariya's

(1985) results, or various generalized UMVU results. The reason is that any attempt to improve the GLS estimator to take account of the constraints results in a biased estimator. In this context, the mean unbiasedness condition is overly restrictive.

On the other hand, estimators $\hat{\delta}$ of $c'\beta_0$ that are median unbiased when no constraints are present can be adjusted quite naturally to take advantage of the restriction that $c'\beta_0$ lies in an interval, or any subset of R, without losing their property of median unbiasedness. Whenever $\hat{\delta}$ lies outside the parameter space of $c'\beta_0$ set the adjusted estimator $(\hat{\delta})_R$ equal to any closest value in the closure of the parameter space; otherwise leave the estimator as is. The resultant estimator $(\hat{\delta})_R$ is median unbiased for the restricted parameter space, and lies in its closure. Thus, the condition of median unbiasedness is a relatively attractive condition for restricting the class of estimators when the parameter space of $c'\beta_0$ is restricted.

We now define the linear combinations $(c_2'\beta_0, \ldots, c_K'\beta_0)$ of β_0 that may be subject to additional restrictions beyond that on $c'\beta_0$. Let $\dot{X} = \Sigma^{-1/2}X$. Since $\dot{X}$ is full rank K, c' is proportional to some linear combination of the rows of $\dot{X}$. Say, $c' \propto d_1'\dot{X}$, where d_1 is an orthonormal N-vector. Take any K-1 orthonormal N-vectors $d_2, \ldots, d_K$ that are orthogonal to d_1 and are such that $(d_1, \ldots, d_K)$ span the column space of $\dot{X}$. Then, the vectors c_j, $j = 2, \ldots, K$ are given by $c_j = d_j'\dot{X}$, for $j = 2, \ldots, K$. As a simple example, suppose $\Sigma = I_N$, $c' = (0, \ldots, 0, 1)$, so that $c'\beta_0 = \beta_{0K}$, and the K^{th} column of X is orthogonal to its other columns. In this case, $(\beta_{01}, \ldots, \beta_{0K-1})$ can be restricted in any way (not involving β_{0K}) without affecting the optimality of the best median unbiased estimator of β_{0K}.

The main result of this section gives a strong optimality property for the restricted GLS estimator $(c'\hat{\hat{\beta}})_R$:

Theorem 2. *Consider the model* $y = X\beta_0 + u_N$, *when* $\mathcal{L}(u_N) \in \mathcal{G}_N$, X *has full rank,* $c'\beta_0$ *lies in a known (possibly infinite) interval* I_c *that does not depend on* β_0 , *and the linear combinations* $(c_2'\beta_0, \ldots, c_K'\beta_0)$ *of* β_0 *are restricted in any fashion not involving* $c'\beta_0$. *Then, the restricted GLS estimator* $(c'\hat{\hat{\beta}})_R$ *is the unique best median unbiased estimator of* $c'\beta_0$ *in the sense of uniformly minimum risk for any monotone loss function. Equivalently, it is the unique stochastically best median unbiased estimator of* $c'\beta_0$, *for given* $c \in R^K$.

The proof of this result makes use of the Theorem of Pfanzagl (1979), see Section 5.

Comment: When the restrictions on β_0 are such that the interval containing $c'\beta_0$ depends on β_0 , a uniformly best median unbiased estimator of $c'\beta_0$ does not exist. We still can obtain a lower bound on the risk of a median unbiased estimator of $c'\beta_0$, however, by using the method of the proof of Theorem 2. In particular, if we suppose $(c_2'\beta_0, \ldots, c_K'\beta_0)$ are known, then the interval containing $c'\beta_0$ is known, call it $I_c(\beta_0)$, and the stochastically best median unbiased estimator of $c'\beta_0$ is the restricted GLS estimator $(c'\hat{\hat{\beta}})_R$, restricted to the closure of $I_c(\beta_0)$. The risk of $(c'\hat{\hat{\beta}})_R$ as a function of β_0 gives the desired lower bound.

4. OPTIMAL ESTIMATION OF σ^2

In this section we specialize to the case of linear regression with independent identically distributed normal errors with mean zero and variance σ^2. We consider estimation of σ^2 and various monotone transformations of σ^2, such as σ and $d\sigma^2$, for some constant $d \neq 0$.

First, we discuss the optimality properties of the most commonly used estimators, viz., s^2, s, and ds^2, for σ^2, σ, and $d\sigma^2$, respectively, where $s^2 = \frac{1}{N-K}(y - X\hat{\beta})'(y - X\hat{\beta}) \equiv \frac{1}{N-K}SSR$ and $\hat{\beta}$ is the least squares estimator. The use of s^2 to estimate σ^2 is justified in this context by the fact that it is the best unbiased estimator in the sense of uniformly minimum risk for any convex loss function (see Lehmann (1983, Theorem 3.4.1, p. 185)). This optimality property carries over the estimation of $d\sigma^2$ by ds^2, but does not hold for the standard error of estimate (SEE) s of σ, since s is biased. If σ^2 is known to lie in a non-degenerate interval strictly contained in R^+, then s^2 is still the best unbiased estimator of σ^2 for convex loss, even though it ignores the restrictions on σ^2.

For squared error loss, the risk of s^2 is uniformly dominated by that of the biased estimator $\tilde{s}^2 = \frac{1}{N-K+2}SSR$ (e.g., see Rao (1973, p. 316)). This result is not of great concern, however, since the *symmetric* squared error loss function is usually quite inappropriate for estimation of σ^2. For example, it implies that the maximum loss from under-estimation is bounded whereas that from over-estimation is unbounded. Furthermore, by appropriate choice of asymmetric squared error loss function, s^2 dominates $\tilde{s}^2$ and any other scalar multiple of SSR.

We now consider an alternative to s^2 and $h(s^2)$ for estimating σ^2

and $h(\sigma^2)$, where $h(\cdot)$ is any monotone function. This alternative has several desirable properties. Suppose σ^2 is known to lie in an interval with endpoints a , b where $0 \leq a < b \leq \infty$. Define the estimator τ^2 by

$$\tau^2 = \begin{cases} b \\ SSR/m_{N-K} \\ a \end{cases} \text{ when } SSR/m_{N-K} \begin{cases} \geq b \\ \in [a,b] \\ \leq a \ , \end{cases} \tag{8}$$

where m_{N-K} is the median of a chi-square random variable with $N-K$ degrees of freedom.

The estimator τ^2 has the following properties: (i) τ^2 is the best median unbiased estimator of τ^2 for any monotone loss function. Equivalently, it is the stochastically best median unbiased estimator of σ^2 . In contrast to the optimality results for s^2 , the above result includes bounded asymmetric loss functions. (ii) The optimality result of (i) holds even when β_0 is subject to restrictions, provided the parameter space of β_0 has a non-empty interior. (iii) τ^2 take advantage of the restrictions on σ^2 . This is a distinct advantage of τ^2 over the best mean unbiased estimator s^2 . (iv) The estimator $h(\tau^2)$ of $h(\sigma^2)$ inherits the same optimality properties as τ^2 , provided $h(\cdot)$ is monotone on $[a,b]$. In particular, τ and $d\tau^2$ are best median unbiased estimators of σ and $d\sigma^2$, respectively, for any monotone loss function. This result not only guarantees the existence of a best median unbiased estimator for many estimands $h(\sigma^2)$ of interest, it also provides very simple expressions for such estimators. Best mean unbiased estimators of $h(\sigma^2)$ do not exist for some functions $h(\cdot)$, and even when they do exist, they are more difficult to determine than the best median unbiased estimator.

Results (i) and (ii) above follow by showing that the present problem is covered by Pfanzagl's (1979) Theorem, and that the estimator τ^2 is the optimal estimator defined in his proof. Result (iv) follows from (i) and (ii) using the fact that both $h(\cdot)$ and the loss functions under consideration are monotone.

Since it is natural to compare τ^2 and s^2, we might ask: In what ways, and to what extent, do τ^2 and s^2 differ? To answer the first part of this question, we note that $m_{N-K} < N-K$, because m_{N-K} and $N-K$ are the median and mean of a chi-square random variable, respectively. Hence, $s^2 \leq b$ if and only if $s^2 \leq \tau^2$. That is, τ^2 is larger than s^2 unless s^2 takes a value larger than any value in the parameter space of σ^2.

The extent to which τ^2 and s^2 differ depends on two separate factors: (i) whether a is positive and/or b is finite, and if so, on the proximity of the true parameter σ^2 to one or other of the endpoints a or b, and (ii) the size of $N-K$. When $a > 0$ and/or $b < \infty$, τ^2 is a censored or doubly-censored version of SSR/m_{N-K}. The closer is the true value of σ^2 to a or b, the greater is the extent of the censoring.

If $a = 0$ and $b = \infty$, the only difference between τ^2 and s^2 is in the multiplicative constants $1/m_{N-K}$ and $1/(N-K)$. As $N-K \to \infty$, $(N-K)/m_{N-K} = \tau^2/s^2 \to 1$, as expected. For degrees of freedom $N-K$ equal to 10, 20, and 30, m_{N-K} equals 9.342, 19.34, and 29.34, and τ^2 exceeds s^2 by 7.1, 3.6, and 2.4%, respectively. (See Thompson (1941) and Pearson and Hartley (1958, p. 130) for tables giving the medians of chi-square random variables with degrees of freedom less than or equal to one hundred.) τ^2 is equal to s^2 with its degrees of freedom reduced by .66 when $N-K$ is in [8,50) and by .67 when $N-K$ is in [50,100]. More sizeable differences between τ^2 and s^2 occur only if the parameter space is restricted.

5. PROOFS

Proof of Theorem 1. The distribution of $\Sigma^{-1/2}u_N$ can be decomposed into a probability mass at the origin and an absolutely continuous component. Since the GLS estimator equals β_0 if $u_N = \underset{\sim}{0}$, we can assume u_N has no mass at the origin, without loss of generality. Then, $\Sigma^{-1/2}u_N$ has Lebesgue density $\int(2\pi w)^{-N/2}\exp(-\|u_N\|^2/2w)\,dG(w)$, where $G(w)$ is a distribution on $(0,\infty)$ and $\|\cdot\|$ denotes the Euclidean norm on R^N. Let W denote the scalar mixing random variable with distribution $G(\cdot)$. Conditional on $W = w$, the distribution of u_N is multivariate normal with mean $\underset{\sim}{0}$ and covariance $w\Sigma$.

Condition on $W = w$. Let $\tilde{X} = \frac{1}{\sqrt{w}}\Sigma^{-1/2}X$. We can construct an $N \times N$ orthogonal matrix D such that the first K rows of D span the column space of $\tilde{X}$, and the first row of $D\tilde{X}$ is proportional to c'. That is, $d_1'\tilde{X} = \gamma c'$, for the constant $\gamma = \|d_1'\tilde{X}\tilde{X}'d_1\|/\|c\|$, where d_1 denotes the first row of D written as a column. Transform the model by pre-multiplication by $\frac{1}{\gamma\sqrt{w}}D\Sigma^{-1/2}$ to get: $y^* = X^*\beta_0 + u_N^*$, where $y^* = \frac{1}{\gamma\sqrt{w}}D\Sigma^{-1/2}y$, $X^* = \frac{1}{\gamma\sqrt{w}}D\Sigma^{-1/2}X$, and $u_N^* = \frac{1}{\gamma\sqrt{w}}D\Sigma^{-1/2}u_N \sim N\left(\underset{\sim}{0}, \frac{1}{\gamma^2}I_N\right)$. Define $\eta = (\eta_1, \ldots, \eta_N)' = X^*\beta_0$. By the choice of D, we have $(\eta_{K+1}, \ldots, \eta_N) = (0, \ldots, 0)$, and $\eta_1 = e_1'D\tilde{X}\beta_0/\gamma = c'\beta_0$, where $e_1 = (1, 0, \ldots, 0)'$. Thus, the estimand is η_1.

Consider estimation of η_1 when the single observation $y_1^* \sim N(\eta_1, 1/\gamma^2)$ is observed, and $1/\gamma^2$ is assumed known, where y_1^* is the first element of $y^* = (y_1^*, \ldots, y_N^*)'$. The family of densities of y_1^* for $\eta_1 \in R$ forms a monotone likelihood ratio family, and the likelihood ratios are a non-decreasing function of the continuous random variable y_1^*. Hence, by the confidence bound results of Lehmann (1959, Corollary 3, p. 80 and p. 83) for scalar parameters, the unique uniformly minimum risk median unbiased

estimator of η_1 (based on observing y_1^* only) is y_1^*, for any monotone loss function, over the class of non-randomized and randomized estimators. (See Lehmann (1959, p. 81) for construction of the randomized confidence bounds needed to compare the risk of y_1^* with the risks of randomized estimators.)

Now, any unconditionally median unbiased estimator $\hat{\delta}(y,X)$ of $c'\beta_0$ also is median unbiased conditional on $W = w$, because the conditional distribution of u_N is itself a CES distribution. We can write $\hat{\delta}(y,X)$ as $\tilde{\delta}(y^*, X^*)$. For purposes of comparing the risk of $\tilde{\delta}(y^*, X^*)$ with that of y_1^*, suppose the vector $(\eta_2, \ldots, \eta_k)$ is known. The independence of y_1^* and $(y_2^*, \ldots, y_N^*)$, plus the knowledge of X^* and the distribution of $(y_2^*, \ldots, y_N^*)$, implies that $\tilde{\delta}(y^*, X^*)$ has the same conditional distribution as some randomized estimator of η_1 based on the single observation y_1^*. Lehmann's result then implies that conditional on $W = w$, the risk of y_1^* is less than or equal to that of $\tilde{\delta}(y^*, X^*)$. Since the optimal estimator y_1^* does not depend on γ, the assumption of known γ is innocuous. The optimality of y_1^* holds for all w, so integrating out w yields the unconditional optimality of y_1^*. This gives the desired result, because y_1^* is the GLS estimator of $c'\beta_0$: Let $\tilde{y} = \frac{1}{\sqrt{w}} \Sigma^{-1/2} y$, then $y_1^* = \frac{1}{\gamma} d_1' \tilde{y} = \frac{1}{\gamma} d_1' \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} = c'(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} = c'\hat{\beta}$. □

The above proof could be shortened somewhat by applying Pfanzagl's (1979) Theorem, instead of Lehmann's result. This is not done, however, because the proof given above is needed in the proof of Theorem 2 to attain the stated generality of Theorem 2. In addition, the proof of Pfanzagl's result is more complicated than that of Lehmann, because Pfanzagl considers cases where the best estimator is randomized. Thus, the reference to the simpler result of Lehmann may be helpful to the reader.

The extension of Theorem 1 to include non-CES elliptically symmetric distributions (as considered by Kariya (1985) and Hwang (1985)) is problematic using the method of proof given above. Almost all elliptically symmetric distributions can be written as variance mixtures of multivariate normal distributions (see Chu (1973)). For non-CES distributions, however, the mixing "densities" are somewhere negative. The risk inequalities that hold for given variance values are reversed for negative values of the mixing density, and hence, cannot be integrated up over the range of values of the mixing density. Fortunately, as the discussion of Section 2 indicates, the restriction to CES distributions is not serious. The elliptically symmetric distributions of greatest relevance are CES distributions.

Proof of Theorem 2. Proceed as in the proof of Theorem 1 to transform the model such that η_1 is the estimand. The restricted estimator $(y_1^*)_R$ is median unbiased and equals $(c'\hat{\beta})_R$ by arguments given above.

Condition on $W = w$. The linear combinations $(c_2'\beta_0, \ldots, c_K'\beta_0)$ equal $\gamma\sqrt{w}(\eta_2, \ldots, \eta_K)$. Thus, the restrictions on $(c_2'\beta_0, \ldots, c_K'\beta_0)$ do not affect the conditional distribution of y_1^* or the parameter space I_c of η_1. Hence, we can mimic the proof of Theorem 1 and assume w and $(\eta_2, \ldots, \eta_K)$ are known for the purposes of comparing the risk of an arbitrary (conditionally and unconditionally) median unbiased estimator $\tilde{\delta}(y^*, X^*)$ with that of $(y_1^*)_R$. $\tilde{\delta}(y^*, X^*)$ has distribution equal to that of some randomized estimator of η_1 for the case where only y_1^* is observed. Thus, it suffices to show that conditional on $W = w$, $(y_1^*)_R$ is the unique best median unbiased estiamtor of $c'\beta_0$ based on the single observation $y_1^* \sim N(\eta_1, 1/\gamma^2)$. This follows by Pfanzagl's (1979) Theorem. □

Pfanzagl's (1979) Theorem allows for nuisance parameters, and hence, could be applied in the proof of Theorem 2 by treating $(\eta_2, \ldots, \eta_K)$ as nuisance parameters. This approach limits the restrictions on $(\eta_2, \ldots, \eta_K)$, however, because it requires the assumption that the restricted parameter space of $(\eta_2, \ldots, \eta_K)$ contains a non-empty interior.

REFERENCES

Andrews, D. W. K. (1983), "Robust and Efficient Estimation of Nonlinear Regression Models with Dependent Errors," unpublished manuscript, Cowles Foundation for Research in Economics, Yale University.

_____ (1986), "A Note on the Unbiasedness of Feasible GLS, Quasi-Maximum Likelihood, Robust, Adaptive, and Spectral Estimators of the Linear Model," *Econometrica*, forthcoming.

Andrews, D. W. K. and P. C. B. Phillips (1985), "An Extension of Kariya's Version of the Gauss-Markov Theorem for Nonlinear Estimators," unpublished manuscript, Cowles Foundation for Research in Economics, Yale University.

Bennett, B. M. (1961), "On a Certain Multivariate Non-normal Distribution," *Proceedings of the Cambridge Philosophical Society*, 30, 178-191.

Chu, K.-C. (1973), "Estimation and Decision for Linear Systems with Elliptical Random Processes," *IEEE Transactions on Automatic Control*, 18, 499-505.

Dunnett, C. W. and M. Sobel (1955), "Approximations to the Probability Integral and Certain Percentage Points of a Multivariate Analog of Student's t-Distribution," *Biometrika*, 42, 258-260.

Hwang, J. T. (1985), "Universal Domination and Stochastic Domination: Estimation Simultaneously under a Broad Class of Loss Functions," *Annals of Statistics*, 13, 295-314.

Kariya, T. (1985), "A Nonlinear Version of the Gauss-Markov Theorem," *Journal of the American Statistical Association*, 80, 476-477.

Kelker, D. (1970), "Distribution Theory of Spherical Distributions and a Location-Scale Parameter Generalization," *Sankhya*, A32, 419-430.

King, M. L. (1980), "Robust Tests for Spherical Symmetry and Their Application to Least Squares Regression," *Annals of Statistics*, 8, 1265-1271.

Lehmann, E. L. (1959), *Testing Statistical Hypotheses*, New York: Wiley.

_____ (1983), *Theory of Point Estimation*, New York: Wiley.

Lord, R. D. (1954), "The Use of Hankel Transforms in Statistics. I. General Theory and Examples," *Biometrika*, 41, 44-55.

Pearson, E. S. and H. O. Hartley (1958), *Biometrika Tables for Statisticians*, Volume 1, Second Edition, London: Cambridge University Press.

Pfanzagl, J. (1979), "On Optimal Median Unbiased Estimators in the Presence of Nuisance Parameters," *Annals of Statistics*, 7, 187-193.

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

Schoenberg, I. J. (1938), "Metric Spaces and Completely Monotone Functions," *Annals of Mathematics*, 39, 811-841.

Thompson, Catherine M. (1941), "Tables of the Percentage Points of the χ^2-distribution," *Biometrika*, 32, 187-191.

Widder, D. V. (1941), *The Laplace Transform*, Princeton: Princeton University Press.