LEAST SQUARES AND BEST UNBIASED ESTIMATES[1]

By T. W. Anderson

Columbia University

1. Introduction. The Gauss-Markov Theorem states that least squares estimates are best linear unbiased estimates. A probability model for the assertion specifies that each observable variable can be written

$$y_\alpha = \sum_{i=1}^{p} \beta_i x_{i\alpha} + v_\alpha , \qquad \alpha = 1, \cdots, n, \tag{1}$$

where $\beta_1, \cdots, \beta_p$ are parameters to be estimated, the set $x_{i\alpha}$ are known numbers, forming a matrix of rank $p(\leqq n)$ and $v_1, \cdots, v_n$ are (unobservable) random variables with means 0, variances σ^2 and are uncorrelated. *Best* means minimum variance among unbiased estimates. In this paper we raise the question of the extent to which the qualification *linear* can be omitted from the statement of the theorem.

We shall assume that the errors in (1), $v_1, \cdots, v_n$, are independently distributed with means 0 and common variances σ^2 and in the first part of the note that they are identically distributed. Then *unbiased* means unbiased identically in the values of $\beta_1, \cdots, \beta_p$ and the common error distribution; *minimum variance* means uniformly with respect to these parameters and the distribution. We consider estimates of every nontrivial linear combination $\sum_1^p \theta_i\beta_i$. The least squares estimate of the linear combination is $\sum_1^p \theta_i b_i$, where

$$(b_1, \cdots, b_p) = b' = \sum_1^n y_\alpha x'_\alpha \Big(\sum_1^n x_\alpha x'_\alpha\Big)^{-1}$$

and $x'_\alpha = (x_{1\alpha}, \cdots, x_{p\alpha})$. For convenience we assume throughout the paper that the rank of the matrix $(x_{i\alpha})$ is p.

2. Case of identically distributed errors. A particular linear combination of the parameters is $\sum_1^p \beta_i\bar{x}_i = \mu$, say, where $\bar{x}_i = \sum_1^n x_{i\alpha}/n$; this is the expected value of the sample mean $\bar{y} = \sum_1^n y_\alpha/n$. In general a least squares estimate is a linear combination of the observations, say $\sum_1^n c_\alpha y_\alpha$, and each coefficient is a linear combination of the corresponding "independent" variates, say $c_\alpha = \sum_1^p \phi_i x_{i\alpha}$. The sample mean $\bar{y}$ is the least squares estimate of μ if there exist p numbers, $\phi_1, \cdots, \phi_p$, such that $1/n = \sum_1^p \phi_i\bar{x}_i$. Then the regression function $\mathcal{E}y_\alpha$ can be written $\mu + \sum_1^{p-1} \eta_i w_{i\alpha}$, where $(1, w_{1\alpha}, \cdots, w_{p-1,\alpha})$ is a linear transform of $(x_{1\alpha}, \cdots, x_{p\alpha})$ and $(\mu, \eta_1, \cdots, \eta_{p-1})$ is the inverse linear transform of $(\beta_1, \cdots, \beta_p)$.

Proposition 1. *If $\bar{y}$ is a least squares estimate, it is the best unbiased estimate of $\sum_1^p \beta_i\bar{x}_i$.*

Received July 16, 1961.

[1] This research was sponsored by the Office of Naval Research under Contract Number Nonr-266(33), Project Number NR 042-034, and Contract Number Nonr-3279(00). Reproduction in whole or in part is permitted for any purpose of the United States Government.

PROOF. Halmos [2] proved the proposition in the special case that the y's are identically distributed (that is, $p = 1, x_{1\alpha} = 1$); the proof of the more general proposition here is different because of the lack of complete symmetry. Let an arbitrary estimate be $\bar{y} + h(y_1, \cdots, y_n)$. This estimate is an unbiased estimate of $\mu = \sum_1^p \beta_i \bar{x}_i$ if

$$\mathcal{E}h(y_1, \cdots, y_n) \equiv 0 \tag{2}$$

since $\mathcal{E}\bar{y} \equiv \mu$. We shall show that (2) holding for all distributions implies

$$\mathcal{E}\bar{y}h(y_1, \cdots, y_n) = 0. \tag{3}$$

Hence, the variance of the estimate is

$$\mathcal{E}[\bar{y} - \mu + h(y_1, \cdots, y_n)]^2 = \mathcal{E}(\bar{y} - \mu)^2 + \mathcal{E}h^2(y_1, \cdots, y_n); \tag{4}$$

this is at least equal to the variance of the least squares estimate, which is the first term on the right.

Let

$$\begin{aligned} h(y_1, \cdots, y_n) &= h(\eta' w_1 + u_1, \cdots, \eta' w_n + u_n) \\ &= g(u_1, \cdots, u_n; \eta), \end{aligned} \tag{5}$$

where $\eta' = (\eta_1, \cdots, \eta_{p-1})$, $w_\alpha' = (w_{1\alpha}, \cdots, w_{p-1,\alpha})$, and $u_\alpha = \mu + v_\alpha$. In order to demonstrate (3) we prove the following lemma.

LEMMA. (2) *implies*

$$\sum_{\text{all perm.}} g(a_{\alpha_1}, \cdots, a_{\alpha_n}; \eta) = 0 \tag{6}$$

for every η and the sum is over all permutations of any n numbers $(a_1, \cdots, a_n)$.

PROOF OF LEMMA. We prove (6) by considering special distributions of u_γ. If $\Pr\{u_\gamma = a_\alpha\} = 1, \gamma = 1, \cdots, n$, (2) is

$$g(a_\alpha, a_\alpha, \cdots, a_\alpha; \eta) = 0. \tag{7}$$

If $\Pr\{u_\gamma = a_{\alpha_k}\} = 1/m, k = 1, \cdots, m$, for a set $(a_{\alpha_1}, \cdots, a_{\alpha_m})$ of $(a_1, \cdots, a_n)$, then

$$0 = m^n \mathcal{E}g(u_1, \cdots, u_n; \eta) = \sum g(z_1, \cdots, z_n; \eta), \tag{8}$$

where $\sum$ denotes the sum in which each z_γ takes on all of the values $a_{\alpha_1}, \cdots, a_{\alpha_m}$ $(m \leqq n)$. Now we prove by induction that the sum of $g(z_1, \cdots, z_n; \eta)$ vanishes when the sum is over only those terms for each of which the arguments of $g(z_1, \cdots, z_n; \eta)$ include all m numbers. This holds for $m = 1$ as indicated in (7). Now we assume the property for $m = 1, 2, \cdots, M - 1$ and prove it for $M(\leqq n)$. The sum (8) is

$$\begin{aligned} &\sum_{k=1}^{M} g(a_{\alpha_k}, a_{\alpha_k}, \cdots, a_{\alpha_k}; \eta) + \sum \sum\nolimits_2 g(z_1, \cdots, z_n; \eta) \\ &\quad + \cdots + \sum \sum\nolimits_{M-1} g(z_1, \cdots, z_n; \eta) + \sum\nolimits_M g(z_1, \cdots, z_n; \eta) = 0, \end{aligned} \tag{9}$$

where $\sum_m$ indicates the sum of all terms involving all m of a subset of m of $a_{\alpha_1}, \cdots, a_{\alpha_M}$ and each of these is summed over all such possible subsets. Each $\sum_m$ is 0 by assumption ($m = 1, \cdots, M-1$), and hence the sum $\sum_M$ is 0. The lemma follows for $M = n$.

Returning to the proof of Proposition 1, we have

$$\mathcal{E}\bar{y}h(y_1, \cdots, y_n) = \frac{1}{n}\mathcal{E}\sum_{\alpha=1}^{n} u_\alpha g(u_1, \cdots, u_n; \eta). \tag{10}$$

Since the u's are identically and independently distributed, we can relabel the u's in $\sum u_\alpha g(u_1, \cdots, u_n; \eta)$ to obtain equivalently $\sum u_\alpha g(u_{\alpha_1}, \cdots, u_{\alpha_n}; \eta)$, where $(\alpha_1, \cdots, \alpha_n)$ is a permutation of $(1, \cdots, n)$. Thus

$$n!\mathcal{E}\sum_{\alpha=1}^{n} u_\alpha g(u_1, \cdots, u_n; \eta) = \mathcal{E}\sum_{\alpha=1}^{n} u_\alpha \sum_{\text{all perm.}} g(u_{\alpha_1}, \cdots, u_{\alpha_n}; \eta). \tag{11}$$

The lemma implies (11) is 0, and this fact implies Proposition 1. (Note we did not use the rank condition on the independent variables or that $n \geqq p$.)

PROPOSITION 2. *If $n = p$, every least squares estimate $\sum_1^p \theta_i b_i$ is the best unbiased estimate of $\sum_1^p \theta_i\beta_i$, where b_i is the least squares estimate of β_i.*

PROOF. An arbitrary unbiased estimate is $\sum \theta_i b_i + h(y_1, \cdots, y_n)$ with (2) holding. When $v_i = 0$,

$$h\left(\sum_{i=1}^{p}\beta_i x_{i1}, \cdots, \sum_{i=1}^{p}\beta_i x_{ip}\right) = 0 \tag{12}$$

for every set $\beta_1, \cdots, \beta_p$. Since the arguments of h are a nonsingular linear transformation of $\beta_1, \cdots, \beta_p$, $h(y_1, \cdots, y_p) = 0$ for every set of arguments. This shows that $\sum \theta_i b_i$ is the *only* unbiased estimate of $\sum \theta_i\beta_i$ and hence is the best.

PROPOSITION 3. *If the deletion of each column from the matrix $(x_1, \cdots, x_n)$ leaves a matrix of rank p no least squares estimate not proportional to $\bar{y}$ is a best unbiased estimate.*

PROOF. We demonstrate the proposition by displaying an unbiased estimate that for some distribution has a smaller variance than $\sum \theta_i b_i = \sum c_\alpha y_\alpha$. Let the estimate to be determined be of the form $\sum c_\alpha y_\alpha + h(y_1, \cdots, y_n)$, where

$$h(y_1, \cdots, y_n) = z\sum_{\alpha,\gamma=1}^{n} a_{\alpha\gamma} y_\alpha y_\gamma, \tag{13}$$

where $(a_{\alpha\gamma})$ is a symmetric matrix and z is a scalar. The condition of unbiasedness is

$$\begin{aligned} 0 = \mathcal{E}h(y_1, \cdots, y_n) &= z\sum_{\alpha,\gamma=1}^{n} a_{\alpha\gamma}\mathcal{E}y_\alpha y_\gamma \\ &= z\left[\sigma^2\sum_{\alpha=1}^{n} a_{\alpha\alpha} + \sum_{i,j=1}^{p}\beta_i\beta_j\sum_{\alpha,\gamma=1}^{n} a_{\alpha\gamma}x_{i\alpha}x_{j\gamma}\right]. \end{aligned} \tag{14}$$

Since this is an identity in $\sigma^2, \beta_1, \cdots, \beta_p$, the coefficients in (14) of σ^2 and $\beta_i\beta_j$ $(i, j = 1, \cdots, p)$ are all 0. The variance of the estimate is

$$\mathcal{E}\left[\sum_{\alpha=1}^{n} c_\alpha\left(y_\alpha - \sum_{i=1}^{p}\beta_i x_{i\alpha}\right) + z\sum_{\alpha,\gamma=1}^{n} a_{\alpha\gamma}y_\alpha y_\gamma\right]^2$$

$$= \mathcal{E}\left[\sum_{\alpha=1}^{n} c_\alpha\left(y_\alpha - \sum_{i=1}^{p}\beta_i x_{i\alpha}\right)\right]^2 \tag{15}$$

$$+ 2z\mathcal{E}\sum_{\delta=1}^{n} c_\delta\left(y_\delta - \sum_{i=1}^{p}\beta_i x_{i\delta}\right)\sum_{\alpha,\gamma=1}^{n} a_{\alpha\gamma}y_\alpha y_\gamma + z^2\mathcal{E}\left[\sum_{\alpha,\gamma=1}^{p} a_{\alpha\gamma}y_\alpha y_\gamma\right]^2.$$

The second term on the right hand side of (15) is

$$2z\mathcal{E}\sum_{\delta=1}^{n} c_\delta v_\delta \sum_{\alpha,\gamma=1}^{n} a_{\alpha\gamma}\left(v_\alpha + \sum_{i=1}^{p}\beta_i x_{i\alpha}\right)\left(v_\gamma + \sum_{j=1}^{p}\beta_j x_{j\gamma}\right)$$

$$= 2z\left\{\nu_3\sum_{\delta=1}^{n} c_\delta a_{\delta\delta} + 2\sigma^2\sum_{i=1}^{p}\beta_i\sum_{\delta,\gamma=1}^{n} c_\delta a_{\delta\gamma}x_{i\gamma}\right\} = 2z\nu_3\sum_{\delta=1}^{n} d_\delta a_{\delta\delta}, \tag{16}$$

where $\nu_3 = \mathcal{E}v_\alpha^3$. The sums bilinear in c_δ and $x_{i\gamma}$ are 0 because $c_\delta = \sum_i \phi_i x_{i\delta}$ and the identity (14) implies conditions on $(a_{\alpha\gamma}) = A$, which are

$$\sum_{\alpha=1}^{n} a_{\alpha\alpha} = 0, \tag{17}$$

$$\sum_{\alpha,\gamma=1}^{n} a_{\alpha\gamma}x_{i\alpha}x_{j\gamma} = 0, \qquad i, j = 1, \cdots, p. \tag{18}$$

If we can take A so

$$\sum_{\delta=1}^{n} c_\delta a_{\delta\delta} \neq 0, \tag{19}$$

the coefficient of z in (16) is different from 0, and z can be taken so this second term is negative and larger in absolute value than the third term which is quadratic in z. Then the sum of the three terms is less than the first term which is the variance of the least squares estimate.

To complete the proof of the proposition we have to show that there is a set of $\frac{1}{2}n(n+1)$ numbers $a_{\alpha\alpha}$ and $a_{\alpha\gamma} = a_{\gamma\alpha}$ $(\alpha \neq \gamma)$ satisfying (17), (18) and $\sum c_\alpha a_{\alpha\alpha} \neq 0$; that is, that the coefficients of $a_{\alpha\gamma}$ in the inequality are linearly independent of the coefficients in (17) and (18). It may be convenient to think of the $\frac{1}{2}n(n+1)$ coefficients of the different $a_{\alpha\gamma}$ in each linear form of (17), (18) and (19) as constituting a vector; these vectors are linearly independent if there is no linear combination of some vectors equal to another. Thus we shall show that there does not exist a scalar b_0 and a symmetric matrix $B = (b_{ij})$ such that

$$c_\alpha = b_0 + \sum_{i,j=1}^{p} b_{ij}x_{i\alpha}x_{j\alpha}, \tag{20}$$

$$0 = \sum_{i,j=1}^{p} b_{ij}x_{i\alpha}x_{j\gamma}, \qquad \alpha \neq \gamma. \tag{21}$$

The conditions (21) in matrix form are

$$x_\alpha' B x_\gamma = 0, \qquad \alpha \neq \gamma. \tag{22}$$

The conditions of the proposition insure that for each α there are p linearly independent x_γ with $\gamma \neq \alpha$. If (22) holds for each γ, then $x_\alpha' B = 0$. Therefore, $B = 0$, and thus the only solution to (20) is $c_\alpha = b_0$, which defines an estimate proportional to $\bar{y}$.

PROPOSITION 4. *If the deletion of some column from the matrix $(x_1, \cdots, x_n)$ leaves a matrix of rank less than p, any least squares estimate $\sum c_\alpha y_\alpha$ that is a best unbiased estimate involves at most $p + 1$ different weights c_α, and at least $n - p$ weights are equal.*

PROOF. That the least squares estimate $\sum c_\alpha y_\alpha$ is a best unbiased estimate implies that there does not exist an estimate of the form $\sum c_\alpha y_\alpha + h(y_1, \cdots, y_n)$ with h given by (13) satisfying (14) that has a smaller variance. The argument used in proving Proposition 3 shows that in this case there must exist b_0 and B satisfying (20) and (21). Proposition 4 follows from a study of the properties of b_0 and B.

It is convenient to transform the independent variates. Suppose x_1 is a vector whose deletion leaves the matrix of rank less than p and that $(x_1, \cdots, x_p)$ is of rank p. Let C be the matrix such that

$$C(x_1, \cdots, x_p) = I. \tag{23}$$

Let $Cx_\alpha = z_\alpha$ $(\alpha = 1, \cdots, n)$, $\lambda' = \beta' C^{-1}$, $\psi' = \phi' C^{-1}$. Then $\lambda' z_\alpha = \beta' x_\alpha$ and

$$c_\alpha = \sum_{i=1}^{p} \psi_i z_{i\alpha}. \tag{24}$$

Then the hypothesis that $\sum c_\alpha y_\alpha$ is a least squares estimate implies the existence of b_0 and a symmetric matrix $(f_{ij}) = F$ (which is $C'^{-1} B C^{-1}$) such that

$$c_\alpha = b_0 + \sum_{i,j=1}^{p} f_{ij} z_{i\alpha} z_{j\alpha}, \tag{25}$$

$$0 = \sum_{i,j=1}^{p} f_{ij} z_{i\alpha} z_{j\gamma}, \qquad \alpha \neq \gamma. \tag{26}$$

For any pair α, γ $(\alpha \neq \gamma)$, $\alpha, \gamma = 1, \cdots, p$, all the coefficients of f_{ij} in (26) are 0 except that of $f_{\alpha\gamma} = f_{\gamma\alpha}$; hence, $f_{\alpha\gamma} = 0$, $\alpha \neq \gamma$. Thus F is a diagonal matrix. From (24), (25) and the fact that z_i is the ith column of I we deduce

$$c_i = b_0 + f_{ii} = \psi_i, \qquad i = 1, \cdots, p. \tag{27}$$

The conditions (26) are

$$\sum_{i=1}^{p} f_{ii} z_{i\alpha} z_{i\gamma} = 0, \qquad \alpha \neq \gamma, \tag{28}$$

and for $\alpha = i \leqq p$ and $\gamma > p$ the conditions are

$$f_{ii} z_{i\gamma} = 0, \qquad i = 1, \cdots, p, \quad \gamma = p + 1, \cdots, n. \tag{29}$$

Then (25) and the diagonality of F imply that $c_\gamma = b_0$ for $\gamma > p$. This proves the proposition.

Other properties of the matrix F put further restrictions on a least squares estimate that is a best unbiased estimate. We note that

$$\sum_{i=1}^{p} \psi_i z_{i\alpha} = b_0, \qquad \alpha = p+1, \cdots, n. \tag{30}$$

Suppose that exactly r of the rows of $(z_{p+1}, \cdots, z_n)$ are not 0 $(0 \leqq r \leqq p-1)$. (Since $(z_2, \cdots, z_n)$ is of rank less than p, every determinant $|z_\alpha, z_2, \cdots, z_p|$ is 0, $\alpha = p+1, \cdots, n$, which implies $z_{1\alpha} = 0$ because $(z_2, \cdots, z_p)$ are the last $p-1$ columns of I.) Let the components be numbered so these are the last r rows. Then

$$f_{p-r+1,p-r+1} = \cdots = f_{pp} = 0 \tag{31}$$

(from (29) for $i = p-r+1, \cdots, p$), and (27) gives

$$c_i = b_0 = \psi_i, \qquad i = p-r+1, \cdots, p. \tag{32}$$

Then

$$\begin{aligned} b_0 = c_\alpha &= \sum_{i=p-r+1}^{p} \psi_i z_{i\alpha}, \qquad \alpha = p-r+1, \cdots, n, \\ &= b_0 \sum_{i=p-r+1}^{p} z_{i\alpha}. \end{aligned} \tag{33}$$

Hence, either $b_0 = 0$ or

$$1 = \sum_{i=p-r+1}^{p} z_{i\alpha}, \qquad \alpha = p-r+1, \cdots, n. \tag{34}$$

The conditions on A for unbiasedness are

$$\sum_{\alpha,\gamma=1}^{p} a_{\alpha\gamma} z_{i\alpha} z_{j\gamma} = 0, \qquad i, j = 1, \cdots, p. \tag{35}$$

The above discussion implies

$$a_{ij} = 0, \qquad i, j = 1, \cdots, p-r, \tag{36}$$

$$\sum_{\gamma=1}^{n} a_{i\gamma} z_{j\gamma} = 0, \qquad i = 1, \cdots, p-r, j = p-r+1, \cdots, p. \tag{37}$$

If $b_0 = 0$, then $c_\alpha a_{\alpha\alpha} = 0$, $\alpha = 1, \cdots, n$.

The least squares estimates that are best unbiased estimates are quite special as can be seen from the conditions above. There are some other than $\bar{y}$. For example, if $y_1 = \beta_1 + v_1$ and $y_2 = v_2$, then y_1 is the least squares estimate of β_1 and is the best unbiased estimate ($p = 1$, $r = 0$, $b_0 = 0$). If $y_1 = \beta_1 + \beta_2 + v_1$, $y_2 = \beta_1 + \beta_2 + v_2$, $y_3 = \beta_1 - 2\beta_2 + v_3$, then $(y_1 + y_2 - 2y_3)/6$ is the least squares estimate of β_2 and is a best unbiased estimate.

3. Some further remarks. If we do not assume that the errors $v_1, \ldots, v_n$

are identically distributed then, even the least squares estimate $\bar{y}$ is not in general a best unbiased estimate of μ when $n > p$. Consider an unbiased estimate $\sum c_\alpha y_\alpha + h(y_1, \cdots, y_n)$, where $h(y_1, \cdots, y_n)$ is given by (13) and satisfies (14). The variance of this estimate is again (15). The second term on the right of (15) is

$$
\begin{aligned}
&2z\mathcal{E}\sum_{\delta=1}^{n} c_\delta v_\delta \sum_{\alpha,\gamma=1}^{n} a_{\alpha\gamma}(v_\alpha + \beta' x_\alpha)(v_\gamma + \beta' x_\gamma) \\
&\quad = 2z\left(\sum_{\delta=1}^{n} c_\delta a_{\delta\delta}\nu_{3\delta} + 2\sigma^2 \sum_{i=1}^{p} \beta_i \sum_{\alpha,\gamma=1}^{n} c_\alpha a_{\alpha\gamma} x_{i\alpha}\right) = 2z\sum_{\delta=1}^{n} c_\delta a_{\delta\delta}\nu_{3\delta},
\end{aligned} \tag{38}
$$

where $\nu_{3\delta} = \mathcal{E}v_\delta^3$. However, the third-order moments are arbitrary and can be taken so (38) is different from 0 as long as some $c_\delta a_{\delta\delta} \neq 0$, and, in particular, if $c_\delta = 1/n$ and some $a_{\delta\delta} \neq 0$. In fact, the diagonal elements of A can be taken arbitrarily subject to (17), and the nondiagonal elements can be taken to satisfy (18), for the coefficients of $a_{\alpha\gamma} = a_{\gamma\alpha}$, $\alpha \neq \gamma$, in (18) are linearly independent by the proof of Proposition 3 under those assumptions. We summarize these remarks.

PROPOSITION 5. *If $v_1, \cdots, v_n$ are not necessarily identically distributed, and if the deletion of each column from the matrix $(x_1, \cdots, x_n)$ leaves a matrix of rank p, no least squares estimate is a best unbiased estimate of the corresponding parameter.*

It has been shown recently [1] that if the v's are identically and *normally* distributed, any least squares estimate is a best unbiased estimate of the corresponding parameter. This is a consequence of the normality, for $b_1, \cdots, b_p$ are a set of sufficient statistics for $\beta_1, \cdots, \beta_p$ (given σ^2) and the only function of $b_1, \cdots, b_p$ that is an unbiased estimate of a linear combination of β's is that linear combination of b's. This result implies that if there is a best unbiased estimate in general it must be the least squares estimate except on a set of measure 0, but it leaves open the possibility that there is no best unbiased estimate at all because some other unbiased estimate might have smaller variance for some nonnormal distribution.

In this paper all distributions are included. The class of estimates unbiased with respect to all distributions is smaller than the class unbiased with respect to densities; in fact, in the proof of Proposition 1 it is the discrete distributions that are used to characterize the property of the class. However, the property of "best" is stronger for all distributions than for a smaller class because best means uniformly minimum variance. We note that inclusion of discrete distributions permits statements not qualified by exceptions of measure 0.

REFERENCES

[1] A. L. BRUDNO, "On a dispersion proof of the method of least squares" (Russian), *Mat. Sb. N.S.*, Vol. 43 (85) (1957), pp. 37–48. (English translation by Lloyd Rosenberg dittoed in the Dept. of Mathematical Statistics, Columbia University.)

[2] PAUL R. HALMOS, "The theory of unbiased estimation," *Ann. Math. Stat.*, Vol. 17 (1946), pp. 34–43.