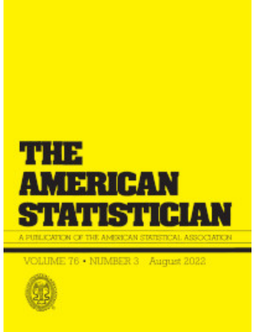

THE AMERICAN STATISTICIAN

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 76 • NUMBER 3 August 2022

Linearity of Unbiased Linear Model Estimators

Stephen Portnoy

THE AMERICAN STATISTICIAN
2022, VOL. 00, NO. 0, 1–4: Short Technical Note
https://doi.org/10.1080/00031305.2022.2076743

Taylor & Francis
Taylor & Francis Group

Linearity of Unbiased Linear Model Estimators

Stephen Portnoy[a,b]

[a]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL; [b]Department of Mathematics and Statistics, Portland State University, Portland, OR

ABSTRACT

Best linear unbiased estimators (BLUE's) are known to be optimal in many respects under normal assumptions. Since variance minimization doesn't depend on normality and unbiasedness is often considered reasonable, many statisticians have felt that BLUE's ought to preform relatively well in some generality. The result here considers the general linear model and shows that any measurable estimator that is unbiased over a moderately large family of distributions must be linear. Thus, imposing unbiasedness cannot offer any improvement over imposing linearity. The problem was suggested by Hansen, who showed that any estimator unbiased for nearly all error distributions (with finite covariance) must have a variance no smaller than that of the best linear estimator in some parametric subfamily. Specifically, the hypothesis of linearity can be dropped from the classical Gauss–Markov Theorem. This might suggest that the best unbiased estimator should provide superior performance, but the result here shows that the best unbiased regression estimator can be no better than the best linear estimator.

ARTICLE HISTORY
Received February 2022
Accepted April 2022

KEYWORDS
Linear estimator; Linear model; Nonparametric; Regression; Unbiased

1. Introductory Comments

Best linear unbiased estimators (BLUE's) are ubiquitous in statistical practice. The fundamental justification for their use comes from normal assumptions, but statisticians now realize that other procedures can show superior performance in certain respects, both inside and outside normal families. Nonetheless, it is often assumed that BLUE's are useful more generally since unbiasedness often seems reasonable. The result here should help to disabuse statisticians of this: for general linear models, any estimator that is unbiased for all distributions in a sufficiently broad family must be linear. Specifically, there is a finite-parameter family of distributions such that unbiasedness over this family implies linearity. Note that second moments are not needed here. First moments are needed to define unbiasedness, but no other requirements are imposed on the estimator (aside from measurability). Further comments on the implications and generality of this result are presented in the Section 3.

The problem was suggested by a recent paper by Hansen (2022) showing that an estimator of a regression parameter that is unbiased over a sufficiently broad family of distributions must achieve the Cramer-Rao lower bound. Since Hansen's family contains the family introduced here, unbiasedness for his family is not more general than linearity. Independently of the work reported here (and contemporaneous with it), Pötscher and Preinerstorfer (2022) prepared a response to Hanson's paper. The authors focused on Gauss–Markov results and so assumed finite second moments. Their paper included a theorem showing that linearity followed from unbiasedness over a class of distributions larger than the class required here. The proof was rather different; and, critically, it assumed finite second moments. The result here requires only first moments and so shows that the implication from unbiasedness to linearity depends neither on higher moments nor on variance optimization. As detailed in Section 3, linearity depends only on unbiasedness over a relatively small (finite-dimensional) set of distributions with finite first moment.

2. Basic Result

The result here considers estimators T of a linear combination $w'\beta$ of regression coefficients. The proof uses (essentially) two-point distributions to compute derivatives of T in each direction and shows the derivative to be constant; from which linearity follows. This requires smoothness of T, which is assumed in Part 1 of the proof. Part 2 uses Lusin's Theorem and somewhat intricate analysis to show that any measurable estimator is an appropriate limit of smooth (and hence, by Part 1, linear) ones. The proofs are given in the Appendix.

Theorem 1. Let $\mathcal{F}$ be a family of (error) distributions that is sufficiently rich so that for any two-point distribution with mean zero, there is a sequence of continuously differentiable densities in $\mathcal{F}$ whose distributions have mean zero, compact support, and converge to a point mass at the origin, and whose convolutions with the two-point distribution are contained in $\mathcal{F}$.

Let $Y = X\beta + e$ where $Y \in \Re^n$, X is a $n \times p$ matrix, $\beta \in \Re^p$, and e has a distribution in $\mathcal{F}$. Let w be any nonzero

CONTACT Stephen Portnoy sportnoy@illinois.edu Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL.

vector in $\mathfrak{R}^p$ .Then any estimator, $T(Y)$, of $w'\beta$ that satisfies $E_F T(Y) = w'\beta$ for all F in $\mathcal{F}$, must be linear almost everywhere (with respect to Lebesgue measure).

Linearity of unbiased estimates of β follows as an immediate consequence: take w_j to be the jth unit co-ordinate vector in $\mathfrak{R}^p$ and apply Theorem 1 to obtain the jth row of a matrix representation for $\underline{T}$. Note that assuming $\underline{T}$ to be an unbiased estimator of the vector β implies that β is estimable and, hence, identifiable. Formally,

Corollary 1. Consider the model and definition of $\mathcal{F}$ above. Then if $\underline{T}(Y)$ is an unbiased estimator of β for all F in $\mathcal{F}$, there is a $p \times n$ matrix, A, such that $\underline{T}(Y) = AY$ almost everywhere.

3. Concluding Remarks

(a) The result here implies Hansen's result (showing the optimality of BLUE's) under conditions requiring unbiasedness over much smaller sets of error distributions. Specifically, it shows that optimality among estimates unbiased over moderately large families of distributions is no stronger than optimality among linear estimators. It may be of interest to explore the relation between nonparametric unbiasedness and invariance, and particularly to see if there is a connection with Eaton and Morris (1970).

(b) Hansen (2022) and Pötscher and Preinerstorfer (2022) both introduce very large classes of distributions over which unbiasedness is required, but their focus on Gauss–Markov theorems has led to requiring second moments. The result here requires only the first moment, which is needed to define unbiasedness. Thus, it clarifies that the linearity result does not depend on consideration of variances (or any other higher moments), nor on considerations of optimality (like the Gauss–Markov Theorem).

(c) The fact that unbiasedness greatly restricts the class of statistical procedures makes it very hard for best unbiased estimators to perform well more generally. For example, in regression settings nonlinear quadratic estimators are often substantially better (see Gnot et al. 1992); not to mention Stein estimation, robust estimation, etc. This appears to be a more general phenomenon. For example, Uniformly Minimum Variance Unbiased (UMVU) estimators are generally found as the unique unbiased estimator depending on a sufficient statistic. Clearly, the best in a class of size one need not be very good, and such estimators can lie outside known bounds and even be quite ridiculous. If X is the number of events observed in the first quarter of a year and is assumed to be Poisson with mean λ, then the UMVU estimator of the probability that there are no more events in the remainder of the year is $T(X) = (-2)^X$. It is often said that there are exceptions to any rule, but the only exceptions where UMVU's possess more general forms of optimality appear to be in the presence of rather strong structural assumptions. For example, in smooth exponential families, the UMVU estimator is asymptotically equivalent to the maximum likelihood estimator (in the sense that the difference is $\mathcal{O}_p(1/n)$, see Portnoy 1977), and so it is asymptotically efficient. However, this requires the rather strong property that in exponential families, the conditional expectation given the mean is $\mathcal{O}(1/n)$ from the unconditional expectation.

(d) It seems rather surprising that linearity requires unbiasedness only over a remarkably small set of distributions. The proof here only uses mixtures of discrete two-point zero-mean distributions with a sequence of distributions with smooth densities that converge in distribution to a point mass at zero. Since two-point distributions form a finite parameter family (pairs of points in $\mathfrak{R}^n$ plus the mixing probability, giving dimension $(2n+1)$), their convolutions with a scale family generates a finite-parameter family of distributions of dimension $(2n+2)$. Thus, fully "nonparametric" (infinite dimensional) unbiasedness is not needed.

(e) While the reference to Lusin (1912) is historically proper, this paper is not readily available and restricts to the real line. Wikipedia has a nice statement of the result for functions on $\mathfrak{R}^n$, as do many modern texts on Measure Theory and Integration; for example see Richardson (2009, sec. 4.4). The result has been generalized to a wide variety of measure spaces.

Acknowledgments

The author thanks Bruce Hansen (whose intriguing article gave rise to this work), Roger Koenker (who brought the article to my attention), Alan Welsh, Benedikt Pötscher, David Preinerstorfer, and the editorial reviewer of the article who offered extensive and very valuable comments. I am very grateful for the helpful remarks and encouragement from these colleagues.

Appendix: Proof of Theorem 1.

The basic idea of the proof is as follows: Part 1 uses unbiasedness of the estimator, $T(Y)$, to relate $T(\pm ay)$ and $T((a+\varepsilon)y)$ in order to compute directional derivatives along a ray. As is clear from the proof in Part 1, this would be straightforward if discrete two-point distributions were allowed. However, discrete distributions will be avoided here since the existence of a gradient is needed to show that linearity of the directional derivatives implies linearity on $\mathfrak{R}^n$. So, Part 1 assumes $T(y)$ is continuously differentiable and considers sequences of (smooth) distributions tending to point masses in order to compute the directional derivatives. The directional derivatives can be expressed in terms of the estimator. This leads to differential equations, which can be solved to show that the directional derivative is constant. The constant must equal the directional derivative at zero, which will imply linearity for differentiable estimators.

Part 2 shows that any estimator unbiased for all $F \in \mathcal{F}$ can be taken to be smooth. This appears to require a somewhat complicated argument and so is done is 4 steps. The first step uses Lusin's Theorem (Lusin 1912: roughly stating that measurable functions are "nearly" continuous), and then obtains a differentiable approximation by using convolution with a sequence of smooth distributions tending to a point mass. Step 2 is a technical proof providing bounds on the unbiased estimator, the smoothed versions, and the contribution from the exceptional set of small measure given by Lusin's Theorem. Step 3 assumes (from Part 1) that the smoothed versions are linear, and uses the previous bounds to show that the linear coefficients converge to a fixed coefficient vector. Finally, Step 4 shows that the smoothed estimators converge to the original unbiased estimator, which will be linear since the linear approximations converge.

In the proof, the set of smoothing distributions are somewhat restricted. They are assumed to have compact support, but they could be given by a fixed scale family (with the scale tending to zero). In fact, mixtures of, say, normals (or t_3-distributions with means but no second moments) will also work, though the lack of a bounded domain complicates the proof. Clearly it is possible to choose a scale family that generates iid observations, but note that the mixture will give stationary but not independent observations.

Proof of Theorem 1

Part 1: Compute directional derivatives in direction y for T smooth:

For Part 1 alone, suppose $T(y)$ has a continuous first derivative. Let $y \in \Re^n$ be a vector with $\|y\| = 1$ and consider error distributions that are the mixture of two distributions concentrated on small neighborhoods around two points, $\{\pm ay\}$. Choose the neighborhoods so that $EY = 0$, and let the neighborhoods shrink. Since T is continuous, the expectations will tend to the (same) mixture of the points. Hence, (by unbiasedness at $w'\beta = 0$)

$$\frac{1}{2}T(ay) + \frac{1}{2}T(-ay) = 0. \tag{1}$$

Now, consider two points $((a+A\epsilon)y)$ and $(-ay)$ with probabilities $(\frac{1}{2} - \epsilon)$ and $(\frac{1}{2} + \epsilon)$ chosen so that the expectation of Y is $\underline{0}$. As above, this implies:

$$(ay + A\epsilon y)(\frac{1}{2} - \epsilon) + (-ay)(\frac{1}{2} + \epsilon) = \underline{0}$$

Simplifying:

$$\frac{1}{2}A\epsilon y - \epsilon(2ay) - A\epsilon^2 y = \underline{0} \quad \Rightarrow \quad A = 4a + \mathcal{O}(\epsilon) \ , \tag{2}$$

where the second equation holds by taking the inner product of the first with y (since $\|y\| = 1$).

Now, as above, unbiasedness of T (at $w'\beta = 0$) implies:

$$T(ay + A\epsilon y)(\frac{1}{2} - \epsilon) + T(-ay)(\frac{1}{2} + \epsilon) = 0. \tag{3}$$

Subtract (1) from (3)

$$\frac{1}{2}(T(y(a + A\epsilon) - T(ay)) + \epsilon(T(-ay) - T(y(a + A\epsilon y)) = 0.$$

Divide by $A\epsilon$ and let $\epsilon \to 0$; note: $A = 4a + \mathcal{O}(\epsilon)$:

$$\partial_y T(ay) = \frac{1}{2a}(T(ay) - T(-ay)). \tag{4}$$

where $\partial_y T(ay)$ denotes the directional derivative of T in direction y (at ay). Similarly:

$$\partial_y T(-ay) = \frac{1}{2a}(T(-ay) - T(ay)) \tag{5}$$

Finally, define

$$S_y^+(a) \equiv T(ay) + T(-ay)$$
$$S_y^-(a) \equiv T(ay) - T(-ay)$$

Then adding and subtracting (4) and (5):

$$\frac{d}{da}\,S_y^+(a) \ = \ 0 \quad ; \quad \frac{d}{da}\,S_y^-(a) \ = \ \frac{1}{a}\,S_y^-(a).$$

Solving the equations for $S_y^+(a)$ and $S_y^-(a)$ (using (1) to show $S_y^+(0) = 0$):

$$S_y^+(a) \ = \ c \ ; \quad S_y^+(0) = 0 \ \Rightarrow \ S_y^+(a) = 0 \ ,$$

$$\frac{d}{da}\,\log(S_y^-(a)) \ = \ \frac{1}{a} \ \Rightarrow \ \log(S_y^-(a)) \ = \ \log(a) \ + \ c_y \quad (a > 0) \ ,$$

where c_y depends on y but is constant in a. Thus,

$$S_y^-(a) \ = \ c_y^*\,a.$$

where $c_y^* = \exp(c_y)$, and again is constant in a (depending on y).

So (for $a > 0$), there is a value $d(y)$, where d is a function not depending on a such that

$$T(ay) \ = \ \frac{1}{2}(S_y^+(a) \ + \ S_y^-(a)) \ = \ 0 \ + \ \frac{1}{2}c_y^*a \ = \ d(y)a \ . \tag{6}$$

Finally, apply (6) in each co-ordinate direction, $\{y_i \ : \ i \ = 1, \ldots, n\}$, where y_i is a fixed unit co-ordinate vector. Thus, $d(y_i)$ is a constant d_i not depending on y_i, and it follows that the gradient of T is the constant vector $b = (d_1, \ldots, d_n)'$; and $T(y) = b'y$.

Part 2. T can be taken to be smooth.

Step 1. Smooth T by convolution:

For each $r > 0$, let $B(r) \subset \Re^n$ denote the ball of radius r. By Lusin's Theorem, one can choose closed sets $E_r \subset B(r)$ so that T restricted to E_r is continuous and $\lambda(B_r - E_r) \leq \varepsilon/2^r$ (where λ is Lebesgue measure). Take $E_\ell = \cup_{r=1:\ell} E_r$ and $E = \lim_\ell E_\ell = \cup_\ell E_\ell$. Then $\{E_\ell\}$ is an increasing sequence of closed sets with $\lambda(\Re^n - E) \leq \varepsilon$, and T restricted to E is continuous.

Let Z_m have a density $g_m(z)$ (with respect to Lebesgue measure), which has an absolutely bounded continuous first derivative and has domain contained in $\{z : \ ||z|| \leq 1/m\}$. Define

$$T_m^*(y) = E_{Z_m} T(y + Z_m) = \int T(y + z)\, g_m(z)\, dz. \tag{7}$$

Then T_m^* is trivially unbiased (since T is, and the distribution of $(e + Z_m)$ is in $\mathcal{F}$), and it has an absolutely bounded and continuous first derivative (as a convolution with such a smooth density). By part 1, $T_m^*(y) = b_m' y$ for some coefficient vector b_m.

Part 2, Step 2. Show T and T_m^* are uniformly bounded on a Ball:

Fix $y \in E$ with $\|y\| \leq r$ (that is, $y \in B(r)$). Since T is continuous on E, and $\|y + z\| \leq r + 1$ for $\|z\| \leq 1$,

$$\sup_{\{(y+z)\in E,\, \|z\|\leq 1\}} |T(y + z)| \ \leq \ \sup_{\{w\in E,\, \|w\|\leq r+1\}} |T(w)| \leq C_r\,, \tag{8}$$

where C_r depends on r, but does not depend on $y \in B(r)$ nor on g_m.

Now, let $I_A(z)$ denote the indicator function of the set A. For each m (and y), choose a set $G_m \subset B(r) \cap (E - y)$, again by Lusin's Theorem for the measure $g_m(z)\,dz$, so that $T(y + z)$ is continuous on G_m and

$$\int I_{G_m^c}(z)\, g_m(z)\, dz \leq \frac{\varepsilon}{1 + \varepsilon} \ , \tag{9}$$

where G_m^c denotes the complement of G_m. From (8),

$$\int I_{G_m^c}(z)\, |T(y + z)|\, g_m(z)\, dz \leq \frac{\varepsilon}{1 + \varepsilon}\, C_r\,, \tag{10}$$

(and C_r remains independent of y for $\|y\| \leq r$, and also of g_m).

Then,

$$T_m^*(y) = E_{Z_m} T(y + Z_m) \tag{11}$$
$$= \int I_{G_m}(z) T(y + z)\, g_m(z)\, dz + \int I_{G_m^c}(z) T(y + z)\, g_m(z)\, dz. \tag{12}$$

The first term needs to be (approximately) an integral with respect to a density; so define

$$d \equiv \int I_{G_m}(z) g_m(z)\, dz = 1 - \int I_{G_m^c}(z) g_m(z)\, dz\,.$$

Then, by (9), $1/(1+\varepsilon) \le d \le 1$, and

$$\left| \int I_{G_m}(z) g_m(z)\, dz - \int I_{G_m}(z)(g_m(z)/d)\, dz \right| \le \left(\frac{1}{d} - 1 \right) \le \varepsilon. \tag{13}$$

Adding (10) and (13) (and using (8)),

$$\left| T_m^*(y) - \int I_{G_m}(z) T(y+z)\, g_m(z)\, dz/d) \right| \le (2C_r + 1)\varepsilon \tag{14}$$

for $y \in E$ with $\|y\| \le r$.

Part 2, Step 3. Show $\{b_m\}$ are uniformly bounded:

Recall that $T_m^*(y) = b_m' y$ (from Part 2) For $\{y_1, \ldots, y_n\}$ in E with $\|y_i\| \le 1$, define $A = A(y_i)$ to be the matrix with rows $\{y_i\}$ and let η_1 be the smallest absolute singular value of A. Consider sets

$$S_a \equiv \{A(y_i) : \ \eta_1 > a,\, y_i \in E,\, \|y_i\| \le 1\}.$$

The complement of S_0 has $\eta_1 = 0$ and so if $A(y_i) \notin S_0$, $\{y_i\}$ must lie in a linear subspace of dimension less than p^2, the dimension of the space of A-matrices. That is, the complement must have measure zero (in $\Re^{p^2}$); and so there must be $a > 0$ such that the measure of S_a is strictly positive (since otherwise the measure of S_0 would be zero). Therefore, there is a matrix of form $A(y_i)$ with $\eta_1 > a$; and, thus, the maximum absolute singular value of A^{-1} is less than the finite constant $1/a$.

Then (with b_m the linear coefficient of T_m^*),

$$(T_m^*(y_i))' = b_m A \ \Rightarrow \ b_m = (T_m^*(y_i))' A^{-1} \Rightarrow \|b_m\| \le \frac{\|(T_m^*(y_i))\|}{a},$$

and from (14),

$$\|b_m\| \le \frac{\sqrt{p}(2C_1 + 1)}{a} \equiv C'. \tag{15}$$

So (by compactness) there is a subsequence along which $\{b_m\}$ converges to b_0. This convergence does not depend on $y \in B(r)$, and so for m large enough, there is a subsequence along which

$$\|b_m - b_0\| \le \varepsilon/r. \tag{16}$$

Part 2, Step 4. Show that T is linear if T_m^* is:

Now, since $T(y+x)$ is continuous on G_m and at y (since $y \in E$),

$$\int I_{G_m}(z) T(y+z)\, (g_m(z)/d)\, dz \ \rightarrow \ T(y).$$

as $m \rightarrow \infty$. This holds for any $y \in E$ with $\|y\| \le r\}$; and this convergence is independent of the convergence of b_m to b_0 shown in Step 3. Therefore, from (16) and (14) (and from $T_m^*(y) = b_m' y$), given $\varepsilon^* > 0$, choose $\varepsilon = \varepsilon^*/C'$ and there is N such that for $m > N$

$$|b_0' y - T(y)| \le |b_m - b_0|\, \|y\| + \left| \int I_{G_m}(z) T(y+z)\, g_m(z)/d\, dz \ - \ T(y) \right| \le 2\varepsilon^*$$

for $y \in E$ with $\|y\| \le r$.

Finally let $\varepsilon^* \rightarrow 0$ and $r \rightarrow \infty$. It follows that $T(y) = b_0' y$ for almost all y (with respect to Lebesgue measure). Q.E.D.

References

Eaton, M., and Morris, C. (1970), "The Application of Invariance to Unbiased Estimation," *The Annals of Mathematical Statistics*, 41, 1708–1716. [2]

Gnot, S., Knautz. H., Trenkler, G., and Zmyslony, R. (1992), "Nonlinear Unbiased Estimation in Linear Models," *Statistics: A Journal of Theoretical and Applied Statistics*, 23, 5-16. [2]

Hansen, B. (2022), "A Modern Gauss–Markov Theorem," *Econometrica*, to appear. [1,2]

Lusin, N. (1912), "Sur les propriétés des fonctions mesurables," *Comptes rendus de l'Académie des Sciences de Paris*, 154, 1688–1690. [2]

Pötscher, B., and Preinerstorfer, D. (2022), "A Modern Gauss-Markov Theorem? Really?," arXiv:2203.01425v2 [math.ST]. [1,2]

Portnoy, S. (1977), "Asymptotic Efficiency of Minimum Variance Unbiased Estimators," *The Annals of Statistics*, 5, 522–529. [2]

Richardson, L. F. (2009), *Measure and Integration: A Concise Introduction to Real Analysis*, Hoboken, NJ: Wiley. [2]