Comment on: "A Modern Gauss-Markov Theorem" by B.E. Hansen[*]

Benedikt M. Pötscher and David Preinerstorfer
Department of Statistics, University of Vienna
SEPS-SEW, University of St. Gallen

First version: May 2022
First revision: December 2022
Second revision: June 2023
Third revision: September 2023

Abstract

We show that Theorem 4 in Hansen (2022) applies to exactly the same class of estimators as does the classical Aitken Theorem. We furthermore point out that Theorems 5-7 in Hansen (2022) contain extra assumptions not present in the classical Gauss-Markov or Aitken Theorem, and thus the former theorems do not contain the latter ones as special cases.

1 Introduction

Hansen (2022) contains several results from which he draws the conclusion that the linearity condition can be dropped from the Aitken Theorem or from the Gauss-Markov Theorem. We argue that this conclusion is unwarranted, as the results on which this conclusion rests either (i) turn out to be equivalent to the classical Aitken or the classical Gauss-Markov Theorem, with linearity being reintroduced indirectly, or (ii) add extra assumptions to the Aitken or Gauss-Markov Theorem.

We thus argue that one should *not* follow Hansen's advice to drop the linearity condition in teaching the Gauss-Markov Theorem or the Aitken Theorem: Depending on which formulation

[*] This note is based on an earlier paper Pötscher and Preinerstorfer (2022) which contains a more extensive discussion, including also a discussion of different versions of Hansen's paper. We thank Abram Kagan for answering an inquiry, Peter Phillips for helpful comments, and the editor Guido Imbens and the referees for their feedback. We are also grateful to Stephen Portnoy for sending us his paper, and to Bruce Hansen for making the page proofs available to us after we had sent him the first version of Pötscher and Preinerstorfer (2022). Address correspondence to Benedikt Pötscher, Department of Statistics, University of Vienna, A-1090 Oskar-Morgenstern Platz 1. E-Mail: benedikt.poetscher@univie.ac.at.

of the Aitken Theorem one starts with (Theorem 3.1 or 3.2 given below), dropping linearity from the formulation of that theorem either leads to a result that is equivalent to the classical Aitken Theorem (if one starts from Theorem 3.2), or leads to an incorrect result (if one starts from Theorem 3.1). The same goes for the Gauss-Markov Theorem.

After the first version of Pötscher and Preinerstorfer (2022), on which the current paper is based, had been circulated, we learned about Portnoy (2022), which establishes, independently and at the same time, a result closely related to our Theorem 3.4 using arguments different from the ones we use; for more discussion see Remark 3.6 in Section 3.

2 The Framework

As in Hansen (2022) we consider throughout the paper the linear regression model

$$Y = X\beta + e \tag{1}$$

where Y is of dimension $n \times 1$ and X is a (non-random) $n \times k$ design matrix with full column rank k satisfying $1 \leq k < n$. It is assumed that

$$Ee = 0 \tag{2}$$

and

$$Eee' = \sigma^2 \Sigma, \tag{3}$$

where $Ee'e < \infty$ ($0 \leq \sigma^2 < \infty$ and Σ a real symmetric nonnegative definite $n \times n$ matrix).[1] While Hansen (2022) does not explicitly assume $\sigma^2 > 0$ and positive definiteness of Σ, both properties are frequently used in his paper. For this reason, we shall in the sequel *always assume* $0 < \sigma^2 < \infty$ and that Σ is a symmetric and positive definite $n \times n$ matrix.

This model implies a distribution F for Y, which, for the given X, depends on β and the distribution of e, in particular on σ^2 and Σ. Now for every Σ define $\mathbf{F}_2(\Sigma)$ as the class of all such distributions F when β varies through $\mathbb{R}^k$ and the distribution of e varies through all distributions compatible with (2) and (3) for the given Σ (and arbitrary σ^2, $0 < \sigma^2 < \infty$). We furthermore introduce the set $\mathbf{F}_2$ as the larger class where we also vary Σ through the set of all symmetric and positive definite $n \times n$ matrices. In other words,

$$\mathbf{F}_2 = \bigcup_{\Sigma} \mathbf{F}_2(\Sigma),$$

where the union is taken over all symmetric and positive definite $n \times n$ matrices.[2] [Of course,

[1] Writing the error covariance matrix as $\sigma^2\Sigma$ is not essential, and we do so only to follow the pertinent literature. Certainly, without a further assumption such as, e.g., 'Σ is known (and nonzero)' the decomposition of Eee' into σ^2 and Σ is not unique.

[2] Note that $\mathbf{F}_2(\Sigma_1) \cap \mathbf{F}_2(\Sigma_2) = \emptyset$ iff Σ_1 and Σ_2 are not proportional. And $\mathbf{F}_2(\Sigma_1) = \mathbf{F}_2(\Sigma_2)$ iff Σ_1 and Σ_2 are proportional.

$\mathbf{F}_2(\Sigma)$ as well as $\mathbf{F}_2$ also depend on the given X, but this dependence is not shown in the notation.] In the following E_F (Var_F, respectively) will denote the expectation (variance-covariance matrix, respectively) taken under the distribution F. A word on notation: Given $F \in \mathbf{F}_2$, there is a unique β, denoted by $\beta(F)$, and a unique $\sigma^2\Sigma$, denoted by $(\sigma^2\Sigma)(F)$, compatible with the distribution F.

3 Aitken and Gauss-Markov Theorems

Let $\hat{\beta}_{GLS} = \hat{\beta}_{GLS}(\Sigma) = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$ denote the generalized least-squares estimator using the matrix Σ (of course, for $\hat{\beta}_{GLS}$ to be feasible, Σ has to be known). Recall that linear estimators are of the form $\hat{\beta} = AY$ where A is a (nonrandom) $k \times n$ matrix. Aitken's Theorem in its usual form (see, e.g., Theil (1971), Section 6.1, Goldberger (1991), Sections 27.1&27.3, Gourieroux and Monfort (1995), Section 6.4.1, Rao and Toutenburg (1995), Theorem 4.4, Hayashi (2000), Proposition 1.7), expressed in the notation of the present paper, reads as follows.

Theorem 3.1. *Let Σ be an arbitrary symmetric and positive definite $n \times n$ matrix. If $\hat{\beta}$ is a linear estimator that is unbiased under all $F \in \mathbf{F}_2(\Sigma)$ (meaning that $E_F\hat{\beta} = \beta(F)$ for every $F \in \mathbf{F}_2(\Sigma)$), then*

$$Var_F(\hat{\beta}) \succeq Var_F(\hat{\beta}_{GLS}(\Sigma))$$

for every $F \in \mathbf{F}_2(\Sigma)$. [Here $\succeq$ denotes Loewner order, i.e., for symmetric matrices Ω_1 and Ω_2 of the same dimension, $\Omega_1 \succeq \Omega_2$ signifies nonnegative definiteness of $\Omega_1 - \Omega_2$.]

The theorem can alternatively be reformulated in the following way.

Theorem 3.2. *Let Σ be an arbitrary symmetric and positive definite $n \times n$ matrix. If $\hat{\beta}$ is a linear estimator that is unbiased under all $F \in \mathbf{F}_2$ (meaning that $E_F\hat{\beta} = \beta(F)$ for every $F \in \mathbf{F}_2$), then*

$$Var_F(\hat{\beta}) \succeq Var_F(\hat{\beta}_{GLS}(\Sigma)) \tag{4}$$

for every $F \in \mathbf{F}_2(\Sigma)$.

In the latter theorem the unbiasedness is requested to hold over the *larger* class $\mathbf{F}_2$ of distributions rather than only over $\mathbf{F}_2(\Sigma)$. Of course, this is immaterial here and the two theorems are equivalent, because the estimators are required to be *linear* in both theorems and thus their expectations depend only on the first moment of Y and not on the second moments at all.

We note that the preceding theorem is equivalent to Theorem 3 in Hansen (2022). To see the equivalence, note that the (implicit) all-quantor over Σ in Theorem 3.2 can be "absorbed" by replacing $\mathbf{F}_2(\Sigma)$ in that theorem with $\mathbf{F}_2$, provided the quantity $\sigma^2\Sigma$ appearing in the expression $Var_F(\hat{\beta}_{GLS}(\Sigma)) = \sigma^2(X'\Sigma^{-1}X)^{-1} = (X'(\sigma^2\Sigma)^{-1}X)^{-1}$ in (4) above is understood as $(\sigma^2\Sigma)(F)$. [Such an understanding is necessary in any case for Theorem 3 in Hansen (2022) to represent a mathematically well-defined statement: Observe that the product $\sigma^2\Sigma$, on which the r.h.s. of the inequality in Theorem 3 in Hansen (2022) depends (note that σ^2 and Σ enter the expression only via the product), is unspecified, and needs to be interpreted as $(\sigma^2\Sigma)(F)$, the variance-covariance matrix of the data under the relevant F w.r.t. which the variance-covariances in this inequality are taken. The same comment applies to Theorems 4 and 5 in Hansen (2022).]

We next discuss what happens if one eliminates the linearity condition in the two equivalent theorems given above. Dropping the linearity conditions leads to the following statements, which will turn out to be *no longer* equivalent to each other:

Statement A: Let Σ be an arbitrary symmetric and positive definite $n \times n$ matrix. If $\hat{\beta}$ is an estimator (i.e., a Borel-measurable function of Y) that is unbiased under all $F \in \mathbf{F}_2(\Sigma)$ (meaning that $E_F\hat{\beta} = \beta(F)$ for every $F \in \mathbf{F}_2(\Sigma)$), then

$$Var_F(\hat{\beta}) \succeq Var_F(\hat{\beta}_{GLS}(\Sigma)) \tag{5}$$

for every $F \in \mathbf{F}_2(\Sigma)$.

Statement B: Let Σ be an arbitrary symmetric and positive definite $n \times n$ matrix. If $\hat{\beta}$ is an estimator that is unbiased under all $F \in \mathbf{F}_2$ (meaning that $E_F\hat{\beta} = \beta(F)$ for every $F \in \mathbf{F}_2$), then

$$Var_F(\hat{\beta}) \succeq Var_F(\hat{\beta}_{GLS}(\Sigma))$$

for every $F \in \mathbf{F}_2(\Sigma)$.

Before proceeding with the discussion of Statements A and B, we need to make a remark on the interpretation of inequalities like (5).

Remark 3.3. (i) In Theorems 3.1 and 3.2 the objects $Var_F(\hat{\beta})$ as well as $Var_F(\hat{\beta}_{GLS}(\Sigma))$ are well-defined as real matrices because all estimators considered are linear, and hence $E_F(\| \hat{\beta} \|^2) < \infty$, $E_F(\| \hat{\beta}_{GLS}(\Sigma) \|^2) < \infty$ holds for every $F \in \mathbf{F}_2(\Sigma)$ where $\| \, . \, \|$ denotes the Euclidean norm. In contrast, in Statements A and B estimators $\hat{\beta}$ with $E_F(\| \hat{\beta} \|^2) = \infty$ for some $F \in \mathbf{F}_2(\Sigma)$ are permissible. [Note that $E_F(\| \hat{\beta} \|^2) = \infty$ for some $F \in \mathbf{F}_2(\Sigma)$ and $E_F(\| \hat{\beta} \|^2) < \infty$ for some other $F \in \mathbf{F}_2(\Sigma)$ may occur.] This necessitates some discussion how Statements A and B are then to be read. For the subsequent discussion note that in both statements $E_F(\hat{\beta})$ is well-defined and finite for every $F \in \mathbf{F}_2(\Sigma)$ as a consequence of the respective unbiasedness assumption (and because $\mathbf{F}_2(\Sigma) \subseteq \mathbf{F}_2$).

(ii) In the *scalar* case (i.e., $k = 1$), there is no problem as the object $Var_F(\hat{\beta})$ is well-defined for every $F \in \mathbf{F}_2(\Sigma)$ as an element of the *extended* real line, regardless of whether $E_F(\| \hat{\beta} \|^2) < \infty$ or not. Hence, inequality (5) always makes sense in case $k = 1$.

(iii) For general k, in case the estimator $\hat{\beta}$ satisfies $E_F(\| \hat{\beta} \|^2) < \infty$ for a given $F \in \mathbf{F}_2(\Sigma)$, the object $Var_F(\hat{\beta})$ is well-defined as a real matrix. Note that the inequality (5) can then equivalently be expressed as $Var_F(c'\hat{\beta}) \geq Var_F(c'\hat{\beta}_{GLS}(\Sigma))$ for every $c \in \mathbb{R}^k$.

(iv) In the case $k > 1$, the object $Var_F(\hat{\beta})$ is not well-defined if $E_F(\| \hat{\beta} \|^2) = \infty$ ($F \in \mathbf{F}_2(\Sigma)$), and hence it is not immediately clear how (5) should then be understood. However, the inequalities $Var_F(c'\hat{\beta}) \geq Var_F(c'\hat{\beta}_{GLS}(\Sigma))$ for every $c \in \mathbb{R}^k$ still make sense in view of (ii) above. We hence may and will interpret (5) (with $F \in \mathbf{F}_2(\Sigma)$) as a *symbolic shorthand notation* for $Var_F(c'\hat{\beta}) \geq Var_F(c'\hat{\beta}_{GLS}(\Sigma))$ for every $c \in \mathbb{R}^k$ (which works both in the case $E_F(\| \hat{\beta} \|^2) < \infty$ and in the case $E_F(\| \hat{\beta} \|^2) = \infty$). We have chosen to write inequality (5) as given (abusing notation), rather than the more conventional and more precise $Var_F(c'\hat{\beta}) \geq Var_F(c'\hat{\beta}_{GLS}(\Sigma))$ for every $c \in \mathbb{R}^k$, in order for our discussion to be easily comparable with the presentation in Hansen's paper, which is silent on this issue.

(v) The above discussion would become moot, if one would introduce the *extra* assumption $E_F(\| \hat{\beta} \|^2) < \infty$ for every $F \in \mathbf{F}_2(\Sigma)$ into Statements A and B. However, such an additional assumption, which has little justification, would (potentially) narrow down the class of estimators competing with $\hat{\beta}_{GLS}(\Sigma)$. As we shall see later on, such an extra assumption actually would have no effect on Statement B (and thus on the corresponding Theorem 4 in Hansen (2022)) at all in view of our Theorem 3.4. The effect it would have on Statement A (and some other results) is discussed in Appendix B of Pötscher and Preinerstorfer (2022).

We now turn to discussing Statements A and B. Not unexpectedly, Statement A is *incorrect* in general.[3] This is known, see, e.g., Gnot et al. (1992), Knautz (1993, 1999), and references therein. For the benefit of the reader we provide a counterexample and attending discussion in Appendix A. In particular, we see that in the Aitken Theorem as it is usually formulated (Theorem 3.1) one can *not* eliminate the linearity condition in general.

Concerning Statement B, observe first that it is equivalent to Theorem 4 in Hansen (2022); this is seen in the same way as the equivalence of Theorem 3.2 above with Theorem 3 in Hansen (2022). A natural question now is why Statement B (i.e., Theorem 4 in Hansen (2022)) would be correct while Statement A is incorrect in general, given that both statements are obtained by dropping one and the same condition (i.e., linearity) from the two equivalent theorems (Theorems 3.1 and 3.2) given above. The answer lies in the fact that Statement B is requiring a *stricter* unbiasedness condition, namely unbiasedness over $\mathbf{F}_2$ rather than only unbiasedness over $\mathbf{F}_2(\Sigma)$. While the two unbiasedness conditions effectively coincide for *linear* estimators as discussed before, this is no longer the case once we leave the realm of linear estimators. Hence, the correctness of Statement B (i.e., of Theorem 4 in Hansen (2022)) crucially rests on imposing the stricter unbiasedness condition, a condition not used in the Aitken Theorem as presented in the references given prior to Theorem 3.1. Note that the class of competitors to $\hat{\beta}_{GLS}(\Sigma)$ figuring in Statement A is, in general, *larger* than the class of competitors appearing in Statement B. Hansen (2022) is quiet on the use of this stricter unbiasedness condition, and no discussion of or motivation for this salient feature of his Theorem 4 is provided.

Having understood what distinguishes Statement B (i.e., Theorem 4 in Hansen (2022)) from (the incorrect) Statement A, the question remains what the scope of the former statement is, i.e.,

[3] I.e., there exist design matrices X such that the statement is false.

how much larger than the class of linear (unbiased) estimators the class of estimators covered by Statement B (i.e., Theorem 4 in Hansen (2022)) really is. We answer this now: As we shall show in the subsequent Theorem 3.4, the only estimators $\hat{\beta}$ satisfying the unbiasedness condition of Statement B (i.e., Theorem 4 in Hansen (2022)) are *linear* estimators. Consequently, Statement B (i.e., Theorem 4 in Hansen (2022)) is equivalent to the Aitken Theorem (i.e., Theorem 3.1 above), as both results give optimality in exactly the same class of estimators.[4] [While the word 'linear' does not appear in the *formulation* of Theorem 4 in Hansen (2022), linearity of the estimators is introduced indirectly through the stricter unbiasedness condition.][5]

We quickly comment on the case where $\Sigma = I_n$. In this case, Theorem 3.1 reduces to the classical Gauss-Markov Theorem, while Theorem 3.2 represents an unusual equivalent reformulation of the Gauss-Markov Theorem. Again, Statement A (with $\Sigma = I_n$) is incorrect in general, see Appendix A. Similar as in the case of general Σ, the correctness of Statement B (with $\Sigma = I_n$) is bought by imposing the stricter unbiasedness condition on the estimators that requires the estimators not only to be unbiased in the model with uncorrelated and homoskedastic errors (which is the model one is studying in the context of the Gauss-Markov Theorem) but also under correlated and/or heteroskedastic errors (i.e., under structures that are 'outside' of the model that is being studied). Why one would want to impose such a requirement seems to be debatable. As already mentioned, the stricter unbiasedness condition employed in Statement B in fact eliminates all nonlinear estimators from consideration (cf. our Theorem 3.4).

What has been said so far also serves as a reminder that one has to be careful with statements such as "best unbiased equals best linear unbiased". While this statement is incorrect in the context of Statement A in general, it is certainly correct in the context of Statement B (i.e., of Theorem 4 in Hansen (2022)) as a consequence of the subsequent Theorem 3.4.

An upshot of the discussion in this section seems to be that – despite an advice to the contrary in Hansen (2022) – one should *not* drop 'linearity' from the pedagogy of the Aitken or Gauss-Markov Theorem: It will lead to an incorrect statement, if one starts from the usual formulation of the classical Aitken Theorem (i.e., from Theorem 3.1); otherwise, if one starts from Theorem 3.2, it will lead to a correct statement which actually is equivalent to the classical Aitken Theorem. The same comment applies to the Gauss-Markov Theorem.

We now provide the theorem alluded to above.

Theorem 3.4. *If $\hat{\beta}$ is an estimator (i.e., a Borel-measurable function of Y) that is unbiased under all $F \in \mathbf{F}_2$ (meaning that $E_F \hat{\beta} = \beta(F)$ for every $F \in \mathbf{F}_2$), then $\hat{\beta}$ is a linear estimator (i.e., $\hat{\beta} = AY$ for some $k \times n$ matrix A).*[6]

Proof: It suffices to establish $\hat{\beta}(y + z) = \hat{\beta}(y) + \hat{\beta}(z)$ as well as $\hat{\beta}(cz) = c\hat{\beta}(z)$ for every y

[4] Recall from before that for linear estimators the unbiasedness conditions in Theorems 3.1 and 3.2 are equivalent.

[5] Adding the extra condition $E_F(\| \hat{\beta} \|^2) < \infty$ for every $F \in \mathbf{F}_2(\Sigma)$ would have no effect on Statement B in view of our Theorem 3.4. The effect this extra condition would have on Statement A is discussed in Appendix B of Pötscher and Preinerstorfer (2022).

[6] By unbiasedness, such an A must then also satisfy $AX = I_k$.

and z in $\mathbb{R}^n$ and every $c \in \mathbb{R}$. For every $m \in \mathbb{N}$ with $m \geq 2$, every $V = (v_1, \ldots, v_m) \in \mathbb{R}^{n \times m}$ and $\alpha \in (0,1)^m$ such that $\sum_{i=1}^m \alpha_i = 1$, define a probability measure (distribution) via

$$\mu_{V,\alpha} := \sum_{i=1}^m \alpha_i \delta_{v_i},$$

where δ_z denotes unit point mass at $z \in \mathbb{R}^n$. The expectation of $\mu_{V,\alpha}$ equals $V\alpha$, and its variance-covariance matrix equals $V \operatorname{diag}(\alpha) V' - (V\alpha)(V\alpha)'$. Denote the expectation operator w.r.t. $\mu_{V,\alpha}$ by $E_{V,\alpha}$. Note that in case $V\alpha = 0$ and $\operatorname{rank}(V) = n$ the measure $\mu_{V,\alpha}$ has expectation zero and a positive definite variance-covariance matrix; thus, $\mu_{V,\alpha}$ corresponds to an $F \in \mathbf{F}_2$ which has $\beta(F) = 0$. From the unbiasedness assumption imposed on $\hat{\beta}$ we obtain that

$$V\alpha = 0 \text{ and } \operatorname{rank}(V) = n \text{ implies } 0 = E_{V,\alpha}(\hat{\beta}) = \sum_{i=1}^m \alpha_i \hat{\beta}(v_i). \tag{6}$$

Step 1: Fix $z \in \mathbb{R}^n$ and define $\alpha^{(1)} = 2^{-1}(n^{-1}, \ldots, n^{-1})' \in \mathbb{R}^{2n}$, $\alpha^{(2)} = 2^{-1}((n+1)^{-1}, \ldots, (n+1)^{-1})' \in \mathbb{R}^{2(n+1)}$, $V_1 = (I_n, -I_n)$ and $V_2 = (I_n, -I_n, z, -z)$. Clearly $V_1\alpha^{(1)} = V_2\alpha^{(2)} = 0$ and $\operatorname{rank}(V_1) = \operatorname{rank}(V_2) = n$. Furthermore,

$$\mu_{V_2,\alpha^{(2)}} = \frac{n}{n+1}\mu_{V_1,\alpha^{(1)}} + \frac{1}{2(n+1)}(\delta_z + \delta_{-z}), \tag{7}$$

which implies

$$E_{V_2,\alpha^{(2)}}(\hat{\beta}) = \frac{n}{n+1}E_{V_1,\alpha^{(1)}}(\hat{\beta}) + \frac{1}{2(n+1)}(\hat{\beta}(z) + \hat{\beta}(-z)).$$

Applying (6) to $E_{V_2,\alpha^{(2)}}(\hat{\beta})$ and $E_{V_1,\alpha^{(1)}}(\hat{\beta})$ now yields $0 = \hat{\beta}(z) + \hat{\beta}(-z)$, i.e., we have shown that

$$\hat{\beta}(-z) = -\hat{\beta}(z) \text{ for every } z \in \mathbb{R}^n, \tag{8}$$

in particular $\hat{\beta}(0) = 0$ follows.

Step 2: Let y and z be elements of $\mathbb{R}^n$. Define the matrix

$$A(y,z) = ((y_1 + z_1)e_1(n), \ldots, (y_n + z_n)e_n(n)),$$

where $e_i(n)$ denotes the i-th standard basis vector in $\mathbb{R}^n$, and set

$$V = (A(y,z), -y, -z, I_n, -I_n) \quad \text{and} \quad \alpha = (3n+2)^{-1}(1, \ldots, 1)' \in \mathbb{R}^{3n+2}.$$

Then, we obtain $V\alpha = 0$ and $\operatorname{rank}(V) = n$. Using (6) and (8) it follows that

$$0 = \sum_{i=1}^n \hat{\beta}((y_i + z_i)e_i(n)) + \hat{\beta}(-y) + \hat{\beta}(-z),$$

which by (8) is equivalent to

$$\hat{\beta}(y) + \hat{\beta}(z) = \sum_{i=1}^{n} \hat{\beta}((y_i + z_i)e_i(n)). \tag{9}$$

Using (9) with y replaced by $y + z$ and z replaced by 0 yields

$$\hat{\beta}(y + z) + \hat{\beta}(0) = \sum_{i=1}^{n} \hat{\beta}((y_i + z_i)e_i(n)).$$

Since $\hat{\beta}(0) = 0$ as shown before, we obtain

$$\hat{\beta}(y) + \hat{\beta}(z) = \hat{\beta}(y + z) \quad \text{for every } y \text{ and } z \text{ in } \mathbb{R}^n. \tag{10}$$

That is, we have shown that $\hat{\beta}$ is additive, i.e., is a group homomorphism between the additive groups $\mathbb{R}^n$ and $\mathbb{R}^k$. By assumption it is also Borel-measurable. It then follows by a result due to Banach and Pettis (e.g., Theorem 2.2 in Rosendal (2009)) that $\hat{\beta}$ is also continuous. Homogeneity of $\hat{\beta}$ now follows from a standard argument, dating back to Cauchy, so that $\hat{\beta}$ is in fact linear. We give the details for the convenience of the reader: Relation (10) (which contains (8) as a special case) implies $\hat{\beta}(lz) = l\hat{\beta}(z)$ for every integer l. Replacing z by z/l ($l \neq 0$) in the latter relation gives $\hat{\beta}(z)/l = \hat{\beta}(z/l)$ for integer $l \neq 0$. It immediately follows that $\hat{\beta}(pz/q) = (p/q)\hat{\beta}(z)$ for every pair of integers p and q ($q \neq 0$). Let $c \in \mathbb{R}$ be arbitrary. Choose a sequence of rational numbers c_s that converges to c. Then by continuity of $\hat{\beta}$

$$\hat{\beta}(cz) = \lim_{s \to \infty} \hat{\beta}(c_s z) = \lim_{s \to \infty} \left(c_s \hat{\beta}(z) \right) = \left(\lim_{s \to \infty} c_s \right) \hat{\beta}(z) = c\hat{\beta}(z).$$

This concludes the proof. ■

Remark 3.5. Inspection of the proof above shows that it does not make use of the full force of the unbiasedness condition ($E_F \hat{\beta} = \beta(F)$ for every $F \in \mathbf{F}_2$), but only exploits unbiasedness for certain strategically chosen discrete distributions F, each with finite support and satisfying $\beta(F) = 0$.

Remark 3.6. Portnoy (2022) uses a somewhat weaker unbiasedness condition than the one used in our Theorem 3.4 (but see Remark 3.5), and then establishes only Lebesgue almost everywhere linearity of the estimators rather than linearity. This is a distinction worth noting for the following reason: The results in Hansen (2022) allow also for discrete distributions. For such distributions positive probability mass can fall into the exceptional Lebesgue null set, showing that any attempt to enforce linearity by appropriately redefining the estimator on the exceptional null set will in general not preserve the statistical properties of the estimator. In particular, the claim in Comment (a) in Section 3 of Portnoy (2022) that his result "implies Hansen's result" is not warranted. Furthermore, at several instances in the discussion in Portnoy (2022) linearity is incorrectly claimed although only linearity Lebesgue almost everywhere is actually established in his paper. For a discussion of other aspects of Portnoy (2022) see Remark 3.6(ii) in Pötscher and Preinerstorfer (2022).

Remark 3.7. In Appendix B we give a "proof" of our Theorem 3.4 above based on Theorem 4.3 in Koopmann (1982) (also reported as Theorem 2.1 in Gnot et al. (1992)), but see the discussion in Appendix B for a caveat.

4 Conclusion

We have shown that the stricter unbiasedness condition employed in Theorem 4 in Hansen (2022) implies linearity of the estimators. It follows that Theorem 4 in Hansen (2022) applies to exactly the same class of estimators as the Aitken Theorem. Thus Theorem 4 in Hansen (2022), although in its formulation new to the literature, is equivalent to the Aitken Theorem. Theorems 5-7 (as well as Theorem 1) in Hansen (2022) are results modelled on the Gauss-Markov or Aitken Theorem but employ extra conditions such as, e.g., independence assumptions. (A more detailed discussion of these results and their scope can be found in Section 5 of Pötscher and Preinerstorfer (2022).) As a consequence, the conclusion drawn in Hansen (2022), that his theorems show that the label "linear estimator" can be dropped from the Gauss-Markov or Aitken Theorem seems to be debatable. We thus repeat our warning against dropping the linearity assumption from the pedagogy of the Gauss-Markov or Aitken Theorem.

A Appendix: Counterexamples

Here we provide a counterexample to Statement A. Further counterexamples can be found in Appendix A of Pötscher and Preinerstorfer (2022). They all rest on the following lemma which certainly is not original as similar computations can be found in the literature, see, e.g., Gnot et al. (1992), Knautz (1993, 1999), and references therein. Counterexamples can also be easily derived from results in the before mentioned papers. In this appendix we always maintain the model from Section 2 and assume that $\Sigma = I_n$ holds. Counterexamples for $\Sigma \neq I_n$ can then easily be obtained by a standard transformation argument. In the following $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$.

Lemma A.1. *Consider the model as in Section 2, additionally satisfying $\Sigma = I_n$.*

(a) Define estimators via

$$\hat{\beta}_{\alpha} = \hat{\beta}_{OLS} + \alpha(Y'H_1Y, \ldots, Y'H_kY)' \tag{11}$$

where the H_i's are symmetric $n \times n$ matrices and α is a real number. Suppose $\operatorname{tr}(H_i) = 0$ *and $X'H_iX = 0$ for $i = 1, \ldots, k$. Then $E_F(\hat{\beta}_{\alpha}) = \beta(F)$ for all $F \in \mathbf{F}_2(I_n)$.*

(b) Suppose the H_i's are as in Part (a). If $Cov_F(c'\hat{\beta}_{OLS}, c'(Y'H_1Y, \ldots, Y'H_kY)') \neq 0$ for some $c \in \mathbb{R}^k$ and for some $F \in \mathbf{F}_2(I_n)$ with finite fourth moments, then there exists an $\alpha \in \mathbb{R}$

such that

$$Var_F(c'\hat{\beta}_\alpha) < Var_F(c'\hat{\beta}_{OLS}); \tag{12}$$

in particular, $\hat{\beta}_{OLS}$ *then does not have smallest variance-covariance matrix (w.r.t. Loewner order) over* $\mathbf{F}_2(I_n)$ *in the class of all estimators that are unbiased under all* $F \in \mathbf{F}_2(I_n)$.[7]

(c) Suppose the H_i*'s are as in Part (a). For every* $c \in \mathbb{R}^k$ *and for every* $F \in \mathbf{F}_2(I_n)$ *(with finite fourth moments) under which* $\beta(F) = 0$ *we have*

$$Cov_F\left(c'\hat{\beta}_{OLS}, c'(Y'H_1Y, \ldots, Y'H_kY)'\right) = \sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{m=1}^{n} d_j \left(\sum_{i=1}^{k} c_i h_{lm}(i)\right) E_F(e_j e_l e_m), \tag{13}$$

where $d = (d_1, \ldots, d_n)' = X(X'X)^{-1}c$ *and* $h_{lm}(i)$ *denotes the* (l,m)*-th element of* H_i.

(d) Suppose the H_i*'s are as in Part (a). For every* $c \in \mathbb{R}^k$ *and for every* $F \in \mathbf{F}_2(I_n)$ *(with finite fourth moments) under which (i)* $\beta(F) = 0$ *and under which (ii) the coordinates of* Y *are independent (equivalently, the errors* e_i *are independent)*

$$Cov_F\left(c'\hat{\beta}_{OLS}, c'(Y'H_1Y, \ldots, Y'H_kY)'\right) = \sum_{j=1}^{n} d_j \left(\sum_{i=1}^{k} c_i h_{jj}(i)\right) E_F(e_j^3). \tag{14}$$

Proof: The proof of Parts (a), (c), and (d) is by straightforward computation. Since

$$\begin{aligned} Var_F(c'\hat{\beta}_\alpha) &= Var_F(c'\hat{\beta}_{OLS}) + 2\alpha Cov_F\left(c'\hat{\beta}_{OLS}, c'(Y'H_1Y, \ldots, Y'H_kY)'\right) \\ &\quad + \alpha^2 Var_F(c'(Y'H_1Y, \ldots, Y'H_kY)'), \end{aligned} \tag{15}$$

the claim in (b) follows immediately as the first derivative of $Var_F(c'\hat{\beta}_\alpha)$ w.r.t. α and evaluated at $\alpha = 0$ equals $2Cov_F\left(c'\hat{\beta}_{OLS}, c'(Y'H_1Y, \ldots, Y'H_kY)'\right)$. Note that all terms in (15) are well-defined and finite because of our fourth moment assumption. Hence, whenever this covariance is non-zero, we may choose $\alpha \neq 0$ small enough such that (12) holds. ■

We now provide a counterexample that makes use of the preceding lemma.

Example A.1. Consider the location model, i.e., the case where $k = 1$ and $X = (1, \ldots, 1)'$. Choose H_1 as the $n \times n$ matrix which has $h_{11}(1) = -h_{22}(1) = 1$ and $h_{ij}(1) = 0$ else. Then the conditions on H_1 in Part (a) of Lemma A.1 are satisfied, and hence $\hat{\beta}_\alpha$ is unbiased under all $F \in \mathbf{F}_2(I_n)$. Setting $c = 1$, we find for the covariance in (14)

$$n^{-1}(E_F(e_1^3) - E_F(e_2^3)) \neq 0$$

for every $F \in \mathbf{F}_2(I_n)$ (with finite fourth moments) under which $\beta(F) = 0$, the errors e_i are independent, and $E_F(e_1^3) \neq E_F(e_2^3)$ hold. Such distributions F obviously exist.[8] As a consequence, $\hat{\beta}_{OLS}$ is not best (over $\mathbf{F}_2(I_n)$) in the class of all estimators $\hat{\beta}$ that are unbiased under

[7] Recall the convention discussed in Remark 3.3.

[8] E.g., choose $e_2, \ldots, e_n$ i.i.d. $N(0, \sigma^2)$ and e_1 independent from $e_2, \ldots, e_n$ with mean zero, variance σ^2, third moment not equal to zero, and finite fourth moment.

all $F \in \mathbf{F}_2(I_n)$. In particular, Statement A (with $\Sigma = I_n$) is false for this design matrix.

For the argument underlying the preceding example it is key that the errors are *not* i.i.d. under the relevant F. In fact, in the location model (i.e., $X = (1, \ldots, 1)'$) we have $Var_F(\hat{\beta}_{OLS}) \preceq Var_F(\hat{\beta}_{\alpha})$ for every real α, for every choice of H_1 as in Part (a) of Lemma A.1, and for every $F \in \mathbf{F}_2(I_n)$ (with finite fourth moments) under which the errors e_i are i.i.d., since then $Cov_F(\hat{\beta}_{OLS}, Y'H_1Y) = 0$ as is easily seen. [This is in line with a result of Halmos (1946) discussed in Section 6 of Pötscher and Preinerstorfer (2022).] For other design matrices X the argument, however, works even for i.i.d. errors as we show in Example A.2 in Pötscher and Preinerstorfer (2022). Cf. Section 4.1 of Gnot et al. (1992) for related results and more.

Many more counterexamples can be generated with the help of Lemma A.1 as discussed in Remark A.2 of Pötscher and Preinerstorfer (2022).

B Appendix: An Alternative "Proof"

We here give a "proof" based on Theorem 4.3 in Koopmann (1982) (also reported as Theorem 2.1 in Gnot et al. (1992)). There is a caveat, however: Theorem 4.3 in Koopmann (1982) is proved by reducing it to Theorem 3.1 (via Theorems 3.2, 4.1, and 4.2) in the same reference. Unfortunately, a full proof of Theorem 3.1 is not provided in Koopmann (1982), only a very rough outline is given. Thus the status of Theorem 4.3 in Koopmann (1982) is not entirely clear. For this reason we have given a direct proof of our Theorem 3.4 in the main text which does not rely on any result in Koopmann (1982).[9]

"Proof": The unbiasedness assumption of the theorem obviously translates into

$$E_F\hat{\beta} = \beta(F) \quad \text{for every} \quad F \in \mathbf{F}_2(\Sigma), \tag{16}$$

for *every* symmetric and positive definite Σ of dimension $n \times n$; specializing to the case $\Sigma = I_n$, we, in particular, obtain[10]

$$E_F\hat{\beta} = \beta(F) \quad \text{for every} \quad F \in \mathbf{F}_2(I_n). \tag{17}$$

Condition (17), together with Theorem 4.3 in Koopmann (1982) (see also Theorem 2.1 in Gnot

[9] Alternatively, one could try to provide a complete proof of the result in Koopmann (1982). We have not pursued this, but have chosen the route via a direct proof of our Theorem 3.4.

[10] Instead of I_n we could have chosen any other symmetric and positive definite $n \times n$ matrix Σ_0 instead.

et al. (1992)[11,12]), implies that $\hat{\beta}$ is of the form

$$\hat{\beta} = A^0 Y + (Y'H_1^0 Y, \dots, Y'H_k^0 Y)', \tag{18}$$

where A^0 satisfies $A^0 X = I_k$ and H_i^0 are matrices satisfying $\operatorname{tr}(H_i^0) = 0$ and $X'H_i^0 X = 0$ for $i = 1, \dots, k$. It is easy to see that we may without loss of generality assume that the matrices H_i^0 are symmetric (otherwise replace H_i^0 by $(H_i^0 + H_i^{0\prime})/2$). Inserting (18) into (16) yields

$$E_F\left(A^0 Y + (Y'H_1^0 Y, \dots, Y'H_k^0 Y)'\right) = \beta(F) \text{ for every } F \in \mathbf{F}_2(\Sigma),$$

and this has to hold for *every* symmetric and positive definite Σ. Standard calculations involving the trace operator and division by σ^2 now give

$$(\operatorname{tr}(H_1^0 \Sigma), \dots, \operatorname{tr}(H_k^0 \Sigma))' = 0 \text{ for every symmetric and positive definite } \Sigma. \tag{19}$$

For every $j = 1, \dots, n$, choose now a sequence of symmetric and positive definite matrices $\Sigma_m^{(j)}$ (each of dimension $n \times n$) that converges to $e_j(n)e_j(n)'$ as $m \to \infty$, where $e_j(n)$ denotes the j-th standard basis vector in $\mathbb{R}^n$ (such sequences obviously exist). Plugging this sequence into (19), letting m go to infinity, and exploiting properties of the trace-operator, we obtain

$$(e_j(n)'H_1^0 e_j(n), \dots, e_j(n)'H_k^0 e_j(n))' = 0 \text{ for every } j = 1, \dots, n.$$

In other words, all the diagonal elements of H_i^0 are zero for every $i = 1, \dots, k$. Next, for every $j, l = 1, \dots, n$, $j \neq l$, choose a sequence of symmetric and positive definite matrices $\Sigma_m^{\{j,l\}}$ (each of dimension $n \times n$) that converges to $(e_j(n) + e_l(n))(e_j(n) + e_l(n))'$ as $m \to \infty$ (such sequences obviously exist). Then exactly the same argument as before delivers

$$((e_j(n) + e_l(n))'H_1^0(e_j(n) + e_l(n)), \dots, (e_j(n) + e_l(n))'H_k^0(e_j(n) + e_l(n)))' = 0 \text{ for every } j \neq l.$$

Recall that the matrices H_i^0 are symmetric. Together with the already established fact that the diagonal elements are all zero, we obtain that also all the off-diagonal elements in any of the matrices H_i^0 are zero; i.e., $H_i^0 = 0$ for every $i = 1, \dots, k$. This completes the proof. ■

Remark B.1. A slightly different version of this "proof" can be obtained as follows. Theorem 4.3 in Koopmann (1982) (together with Footnote 12) shows for every given (fixed) Σ that any $\hat{\beta}$ satisfying (16) is of the form $AY + (Y'H_1 Y, \dots, Y'H_k Y)'$ where $AX = I_k$, the H_i's satisfy $\operatorname{tr}(H_i \Sigma) = 0$, and $X'H_i X = 0$ for $i = 1, \dots, k$. Again it is easy to see that we may assume

[11] Note that X^- in that reference runs through all possible g-inverses of X.

[12] Gnot et al. (1992) assume $\sigma^2 > 0$ whereas Koopmann (1982) allows also $\sigma^2 = 0$. However, both theorems are equivalent as unbiasedness under every $F \in \mathbf{F}_2(I_n)$ also implies unbiasedness under the point distributions at $X\beta$ (i.e., the distributions corresponding to $\sigma^2 = 0$). This is easily seen by considering those distributions in $\mathbf{F}_2(I_n)$ that correspond to $X\beta + e$ with the components of e being independent identically distributed according to $\varepsilon_m(\delta_{-1} + \delta_1)/2 + (1 - \varepsilon_m)\delta_0$. Here ε_m, $0 < \varepsilon_m < 1$, converges to zero for $m \to \infty$ and δ_x denotes point mass at $x \in \mathbb{R}$. A similar argument applies in the case of $\mathbf{F}_2(\Sigma)$.

the matrices H_i to be symmetric. Note that the matrices A and H_i flowing from Theorem 4.3 in Koopmann (1982) in principle could depend on Σ. The following argument shows that this is, however, not the case (after symmetrization of the H_i's) in the present situation: If $\hat{\beta}$ had two distinct linear-quadratic representations with symmetric H_i's, then the difference of these two representations would be a vector of multivariate polynomials (at least one of which is nontrivial) that would have to vanish everywhere, which is impossible since the zero-set of a nontrivial multivariate polynomial is a Lebesgue null-set. Given now the independence (from Σ) of the matrices H_i, one can then exploit the before mentioned relations $\mathrm{tr}(H_i\Sigma) = 0$ in the same way as is done following (19) above.

References

Gnot, S., Knautz, G., Trenkler, G. and Zmyslony, R. (1992). Nonlinear unbiased estimation in linear models. *Statistics*, **23** 5–16.

Goldberger, A. S. (1991). *A course in econometrics.* Harvard University Press, Cambridge, MA.

Gourieroux, C. and Monfort, A. (1995). *Statistics and Econometric Models*, vol. 1. Cambridge University Press.

Halmos, P. R. (1946). The theory of unbiased estimation. *Ann. Math. Statist.*, **17** 34–43.

Hansen, B. E. (2022). A modern Gauss-Markov theorem. *Econometrica*, **90** 1283–1294.

Hayashi, F. (2000). *Econometrics.* Princeton University Press.

Knautz, H. (1993). *Nichtlineare Schätzung des Parameterverktors im linearen Regressionsmodell*, vol. 133 of *Mathematical Systems in Economics.* Verlag Anton Hain, Frankfurt am Main.

Knautz, H. (1999). Nonlinear unbiased estimation in the linear regression model with nonnormal disturbances. *J. Statist. Plann. Inference*, **81** 293–309.

Koopmann, R. (1982). *Parameterschätzung bei a priori Information.* Vandenhoeck & Ruprecht, Göttingen.

Portnoy, S. (2022). Linearity of unbiased linear model estimators. *American Statistician*, **76** 372–375.

Pötscher, B. M. and Preinerstorfer, D. (2022). A modern Gauss-Markov theorem? Really? *arXiv:2203.01425.*

Rao, C. R. and Toutenburg, H. (1995). *Linear models.* Springer Series in Statistics, Springer-Verlag, New York.

Rosendal, C. (2009). Automatic continuity of group homomorphisms. *The Bulletin of Symbolic Logic*, **15** 184–214.

Theil, H. (1971). *Principles of econometrics.* Wiley, New York.