A Modern Gauss-Markov Theorem

Bruce E. Hansen*

University of Wisconsin†

December, 2020
Revised: December 2021

Abstract

This paper presents finite sample efficiency bounds for the core econometric problem of estimation of linear regression coefficients. We show that the classical Gauss-Markov Theorem can be restated omitting the unnatural restriction to linear estimators, without adding any extra conditions. Our results are lower bounds on the variances of unbiased estimators. These lower bounds correspond to the variances of the the least squares estimator and the generalized least squares estimator, depending on the assumption on the error covariances. These results show that we can drop the label "linear estimator" from the pedagogy of the Gauss-Markov Theorem. Instead of referring to these estimators as BLUE, they can legitimately be called BUE (best unbiased estimators).

*Research support from the NSF and the Phipps Chair are gratefully acknowledged. I posthumously thank Gary Chamberlain for encouraging me to study finite sample semi-parametric efficiency, Jack Porter for persuading me to write these results into a paper, and Yuzo Maruyama for catching an error in the proof. I also thank Roger Koenker, Stephen Portnoy, and three referees for thoughtful comments and suggestions.

†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison WI 53706.

1 Introduction

Three central results in core econometric theory are BLUE, Gauss-Markov, and Aitken's. The BLUE theorem states that the best (minimum variance) linear unbiased estimator of a population expectation is the sample mean. The Gauss-Markov theorem states that in a linear homoskedastic regression model the minimum variance linear unbiased estimator of the regression coefficient is the least squares estimator. Aitken's generalization states that in a linear regression model with a general covariance matrix structure the minimum variance linear unbiased estimator is the generalized least squares estimator. These results are straightforward to prove and interpret, and thus are taught in introductory through advanced courses. The theory, however, has a gaping weakness. The restriction to linear estimators is unnatural. There is no justifiable reason for modern econometrics to restrict estimation to linear methods. This leaves open the question if nonlinear estimators could possibly do better than least squares.

One possible answer lies in the theory of uniform minimum variance unbiased (UMVU) estimation (see, e.g., Chapter 2 of Lehmann and Casella (1998)). Lehmann and Casella (1998, Example 4.2) demonstrate that the sample mean is UMVU for the class of distributions having a density. The latter restriction is critical for their demonstration, does not generalize to distributions without densities, and it is unclear if the approach applies to regression models.

A second possible answer is provided by the Cramér-Rao theorem. In the normal regression model the minimum variance unbiased estimator of the regression coefficient is least squares. This result removes the restriction to linearity. But the result is limited to normal regression and so does not provide a complete answer.

A third possible answer is provided by the local asymptotic minimax theorem (see Hajek (1972) and van der Vaart (1998, Chapter 8)) which states that in parametric models, estimation mean squared error cannot be asymptotically smaller than the Cramér-Rao lower bound. This removes the restriction to linear and unbiased estimators, but is focused on a parametric asymptotic framework.

A fourth approach to the problem is semi-parametric asymptotic efficiency, which includes Stein (1956), Levit (1975), Begun, Hall, Huang, and Wellner (1983), Chamberlain (1987), Ritov and Bickel (1990), Newey (1990), Bickel, Klaassen, Ritov, and Wellner (1993), and van der Vaart (1998, Chapter 25). This literature develops asymptotic efficiency bounds for estimation in semi-parametric models including linear regression.

This theory removes the restriction to linear unbiased estimators and parametric models, but only provides asymptotic efficiency bounds, not finite sample bounds. This literature leaves open the possibility that reduced estimation variance might be achieved in finite samples by alternative estimators.

A fifth approach is adaptive efficiency under an independence or symmetry assumption. If the regression error is independent of the regressors and/or symmetrically distributed about zero, efficiency improvements may be possible. If the regression error is fat-tailed, these improvements can be substantial. This literature includes the quantile regression estimator of Koenker and Bassett (1978), the adaptive regression estimator of Bickel (1982), and the generalized t estimator of McDonald and Newey (1988). These improvements are only obtained under the validity of the imposed independence/symmetry assumptions; otherwise the estimators are inconsistent.

Our paper extends the above literatures by providing finite sample variance lower bounds for unbiased estimation of linear regression coefficients without the restriction to linear estimators and without the restriction to parametric models. Our results are semi-parametric, imposing no restrictions on distributions beyond the existence of the first two moments and no restriction on estimators beyond unbiasedness. Our lower bounds generalize the classical BLUE and Gauss-Markov lower bounds, as we show that the same bounds hold in finite samples without the restriction to linear estimators. Our lower bounds also update the asymptotic semi-parametric lower bounds of Chamberlain (1987), as we show that the same bounds hold in finite samples for unbiased estimators.

The results in this paper are a finite-sample version of the insight by Stein (1956) that the supremum of Cramér-Rao bounds over all regular parametric submodels is a lower bound on the asymptotic estimation variance. Our twist turns Stein's insight into a finite-sample argument, thereby constructing a lower bound on the finite-sample variance. Stein's insight lies at the core of semi-parametric efficiency theory. Thus, our result provides a bridge between finite-sample and semi-parametric efficiency theory.

Our primary purpose is to generalize the Gauss-Markov Theorem, providing a finite-sample yet semi-parametric efficiency justification for least squares estimation. A by-product of our result is the observation that it is *impossible* to achieve lower variance than least squares without incurring estimation bias. Consequently, the simultaneous goals of unbiasedness and low variance are incompatible. If estimators are low variance (relative to least squares) they must be biased. This is not an argument against nonparametric, shrinkage, or machine learning estimation, but rather is a statement that these estimation methods should be acknowledged as biased and the latter is necessary to achieve variance reductions.

Our results (similarly to BLUE, Gauss-Markov, Aitken, and Cramér-Rao) focus on unbiased estimators, and thereby are restricted to the special context where unbiased estimators exist. Indeed, the existence of an unbiased estimator is a necessary condition for a finite variance bound. Doss and Sethuraman (1989) showed that when no unbiased estimator exists, then any sequence of estimators with bias tending to zero will have variance tending to infinity. A related literature (Zyskind and Martin (1969), Harville (1981)) concerns conditions for linear estimators to be unbiased when allowing for general covariance matrices.

A caveat is that the class of nonlinear unbiased estimators is small. As shown by Koopmann (1982) and discussed in Gnot, Knautz, Trenkler, and Zmyslony (1992), any unbiased estimator of the regression coefficient can be written as a linear-quadratic function of the dependent variable $\boldsymbol{Y}$. Koopmann's result shows that while nonlinear unbiased estimators exist, they constitute a narrow class.

The literature contains papers which generalize the Gauss-Markov theorem to allow nonlinear estimators, but all are restrictive on the class of allowed nonlinearity, and all are restrictive on the class of allowed error distributions. For example, Kariya (1985) allows for estimators where the nonlinearity can be written in terms of the least squares residuals. Berk and Hwang (1989) and Kariya and Kurata (2002) allow for nonlinear estimators which fall within certain equivariant classes. Each of these papers restricts the error distributions to satisfy a form of spherical symmetry. In contrast, the results presented in this paper do not impose any restrictions on the estimators other than unbiasedness, and do not impose any restrictions on the error distributions.

The proof of our main result (presented in Section 6) is not inherently difficult, but is not elementary either. It might be described as nuanced. It is based on a trick used by Newey (1990, Appendix B) in his development of an asymptotic semi-parametric efficiency bound for estimation of a population expectation.

2 Gauss-Markov Theorem

Let $\boldsymbol{Y}$ be an $n \times 1$ random vector and $\boldsymbol{X}$ an $n \times m$ full-rank regressor matrix with $m < n$. We will treat $\boldsymbol{X}$ as fixed, though all the results apply to random regressors by conditioning on $\boldsymbol{X}$.

The linear regression model is

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

$$\mathbb{E}[\boldsymbol{e}] = 0 \tag{2}$$

$$\operatorname{var}[\boldsymbol{e}] = \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}'\right] = \sigma^2 \boldsymbol{\Sigma} < \infty \tag{3}$$

where $\boldsymbol{e}$ is the $n \times 1$ vector of regression errors. It is assumed that the $n \times n$ matrix $\boldsymbol{\Sigma} > 0$ is known while the scalar $\sigma^2 > 0$ is unknown.

Let $\mathbf{F}_2$ be the set of joint distributions F of random vectors $\boldsymbol{Y}$ satisfying (1)-(3). This is the set of random vectors whose expectation is a linear function of $\boldsymbol{X}$ and has a finite covariance matrix. Equivalently, $\mathbf{F}_2$ consists of all distributions which satisfy a linear regression.

The homoskedastic and serially uncorrelated linear regression model adds the assumption

$$\boldsymbol{\Sigma} = \boldsymbol{I}_n. \tag{4}$$

Let $\mathbf{F}_2^0 \subset \mathbf{F}_2$ be the set of joint distributions satisfying (1)-(4). The standard estimator of β in model $\mathbf{F}_2^0$ is least squares

$$\widehat{\beta}_{\text{ols}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right).$$

For all $F \in \mathbf{F}_2$, $\widehat{\beta}_{\text{ols}}$ is unbiased for β, and for all $F \in \mathbf{F}_2^0$, $\widehat{\beta}_{\text{ols}}$ has variance $\operatorname{var}\left[\widehat{\beta}_{\text{ols}}\right] = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. The question of efficiency is whether there is an alternative unbiased estimator with reduced variance.

The classical Gauss-Markov Theorem applies to *linear* estimators of β, which are estimators that can be written as $\widehat{\beta} = \boldsymbol{A}(\boldsymbol{X})\boldsymbol{Y}$, where $\boldsymbol{A}(\boldsymbol{X})$ is an $m \times n$ function of $\boldsymbol{X}$. Linearity in this context means “linear in $\boldsymbol{Y}$”.

Theorem 1 ***(Gauss-Markov)****. If $\widehat{\beta}$ is a linear estimator, and unbiased for all $F \in \mathbf{F}_2$, then*

$$\operatorname{var}\left[\widehat{\beta}\right] \geq \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

for all $F \in \mathbf{F}_2^0$.

In words, no unbiased linear estimator has a finite sample covariance matrix smaller than the least squares estimator. As this is the exact variance of the least squares estimator, it follows that in the homoskedastic linear regression model, least squares is the minimum variance linear unbiased estimator.

Part of the beauty of the Gauss-Markov Theorem is its simplicity. The only assumptions on the distribution concern the first and second moments of $\boldsymbol{Y}$. The only assumptions on the estimator are linearity and unbiasedness. The statement in the theorem that $\widehat{\beta}$ "is unbiased for all $F \in \mathbf{F}_2$" clarifies the context under which the estimator is required to be unbiased. The requirement that $\widehat{\beta}$ must be unbiased for any distribution means that we are excluding estimators such as $\widehat{\beta} = 0$, which is "unbiased" when the true value satisfies $\beta = 0$. The estimator $\widehat{\beta} = 0$ is not unbiased in the general set of linear regression models $\mathbf{F}_2$ so is not unbiased in the sense of the theorem.

An unsatisfying feature of the Gauss-Markov Theorem is that it restricts attention to linear estimators. This is unnatural as there is no reason to exclude nonlinear estimators. Consequently, when the Gauss-Markov Theorem is taught it is typically followed by the Cramér-Rao Theorem.

Let $\mathbf{F}_2^{\phi} \subset \mathbf{F}_2^0$ be the set of joint distributions satisfying (1)-(4) plus $\boldsymbol{e} \sim \mathrm{N}(0, \boldsymbol{I}_n \sigma^2)$.

Theorem 2 ***(Cramér-Rao)****. If $\widehat{\beta}$ is unbiased for all $F \in \mathbf{F}_2^{\phi}$, then*

$$\operatorname{var}\left[\widehat{\beta}\right] \geq \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

for all $F \in \mathbf{F}_2^{\phi}$.

The Cramér-Rao Theorem shows that the restriction to linear estimators is unnecessary in the class of normal regression models. To obtain this result, in addition to the Gauss-Markov assumptions, the Cramér-Rao Theorem adds the assumption that the observations are independent and normally distributed. The normality assumption is restrictive, however, so neither the Gauss-Markov nor Cramér-Rao Theorem is fully satisfactory. Consequently, the two are typically taught as a pair with the joint goal of justifying the variance lower bound $\sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ and hence least squares estimation.

Closely related to the Gauss-Markov Theorem is the generalization by Aitken (1935) to the context of general covariance matrices. In the linear regression model with nonscalar covariance matrix $\boldsymbol{\Sigma}$, Aitken's generalized least squares (GLS) estimator is

$$\widehat{\beta}_{\mathrm{gls}} = \left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right).$$

For all $F \in \mathbf{F}_2$, $\widehat{\beta}_{\text{gls}}$ is unbiased for β and has variance $\operatorname{var}\left[\widehat{\beta}_{\text{gls}}\right] = \sigma^2 \left(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}\right)^{-1}$. The question of efficiency is whether there is an alternative unbiased estimator with smaller variance. Aitken's Theorem follows Gauss-Markov in restricting attention to linear estimators.

Theorem 3 *(**Aitken**). If $\widehat{\beta}$ is a linear estimator, and unbiased for all $F \in \mathbf{F}_2$, then*

$$\operatorname{var}\left[\widehat{\beta}\right] \geq \sigma^2 \left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}$$

for all $F \in \mathbf{F}_2$.

Aitken's Theorem is less celebrated than the traditional Gauss-Markov Theorem, but perhaps is more illuminating. It shows that, in general, the variance lower bound equals the covariance matrix of the GLS estimator. Thus, in the general linear regression model, generalized least squares is the minimum variance linear unbiased estimator. Aitken's theorem, however, rests on the restriction to linear estimators just as the Gauss-Markov Theorem. In the context of independent observations, Aitken's bound corresponds to the asymptotic semi-parametric efficiency bound established by Chamberlain (1987).

The development of least squares and the Gauss-Markov Theorem involved a series of contributions from some of the most influential probabilists of the nineteenth thru early twentieth centuries. The method of least squares was introduced by Adrien Marie Legendre (1805) as essentially an algorithmic solution to the problem of fitting coefficients when there are more equations than unknowns. This was quickly followed by Carl Friedrich Gauss (1809), who provided a probabilistic foundation. Gauss proposed that the equation errors be treated as random variables, and showed that if their density takes the form we now call "normal" or "Gaussian" then the maximum likelihood estimator of the coefficient equals the least squares estimator. Shortly afterward, Pierre Simon Laplace (1811) justified this choice of density function by showing that his central limit theorem implied that linear estimators are approximately normally distributed in large samples, and that in this context the lowest variance estimator is the least squares estimator. Gauss (1823) synthesized these results and showed that the core result only relies on the first and second moments of the observations and holds in finite samples. Andreĭ Andreevich Markov (1912) provided a textbook treatment of the theorem, and clarified the central role of unbiasedness, which Gauss had only assumed implicitly. Finally, Alexander Aitken (1935) generalized the theorem to cover the case of arbitrary but known covariance matrices. This history, and other details, are documented in Plackett (1949) and Stigler (1986).

3 Modern Gauss-Markov

We now present our main result. We are interested if Aitken's version of the Gauss-Markov Theorem holds without the restriction to linear estimators.

Theorem 4 *If $\widehat{\beta}$ is unbiased for all $F \in \mathbf{F}_2$, then*

$$\operatorname{var}\left[\widehat{\beta}\right] \geq \sigma^2 \left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}$$

for all $F \in \mathbf{F}_2$.

We provide a sketch of the proof in Section 4 and a full proof in Section 6.

Theorem 4 is identical to Theorem 3, but without the limitation to linear estimators. Theorem 4 is a strict improvement, as no additional condition is imposed. This shows that the GLS estimator is the minimum variance unbiased estimator (MVUE) of β.

We can specialize to the context of homoskedastic and serially uncorrelated observations.

Theorem 5 *If $\widehat{\beta}$ is unbiased for all $F \in \mathbf{F}_2$, then*

$$\operatorname{var}\left[\widehat{\beta}\right] \geq \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

for all $F \in \mathbf{F}_2^0$.

Theorem 5 is identical to Theorem 1, but without the limitation to linear estimators. Again, this is a strict improvement. The implication is that in the homoskedastic linear regression model, ordinary least squares is the MVUE of β.

Theorem 5 is also an improvement on Theorem 2 as it lifts the normality assumption of the normal regression model. It is not a strict improvement, however, as the Cramér-Rao Theorem only requires the estimator to be unbiased in the class of normal regression models, while Theorem 5 requires unbiasedness for all regression models.

An important special case of Theorem 5 is estimation of the population expectation. This is the linear regression model where $\boldsymbol{X}$ only contains a vector of ones.

Assume that the elements of $\boldsymbol{Y}$ have a common expectation μ with covariance matrix $\boldsymbol{\Sigma}\sigma^2$. Equivalently, assume $\mathbb{E}[\boldsymbol{Y}] = \mathbf{1}_n\mu$ and $\operatorname{var}[\boldsymbol{Y}] = \boldsymbol{\Sigma}\sigma^2$, where $\mathbf{1}_n$ is a vector of ones. Let $\mathbf{G}_2$ be the set of joint distributions F of random vectors $\boldsymbol{Y}$ satisfying these conditions, and let $\mathbf{G}_2^0$ be the subset with $\boldsymbol{\Sigma} = \boldsymbol{I}_n$. $\mathbf{G}_2^0$ is the set of uncorrelated random variables with a common variance. The standard estimator of μ is the sample mean $\overline{Y}$, which is unbiased and has variance $\operatorname{var}[\overline{Y}] = \sigma^2/n$ for $F \in \mathbf{G}_2^0$.

Theorem 6 *If $\widehat{\mu}$ is unbiased for all $F \in \mathbf{G}_2$, then* $\operatorname{var}\left[\widehat{\mu}\right] \geq \sigma^2/n$ *for all $F \in \mathbf{G}_2^0$.*

As the lower bound σ^2/n equals $\operatorname{var}[\overline{Y}]$, we deduce that the sample mean is the MVUE of μ. Equivalently, the sample mean is the best unbiased estimator (BUE) – there is no need for the classical "linear" modifier.

Essentially, Theorems 4, 5, and 6 show that we can drop the label "linear estimator" from the pedagogy of the Gauss-Markov Theorem. Instead, GLS, OLS, and sample means are the best unbiased estimators of their population counterparts.

4 A Sketch of the Proof

In this section we give an simplified proof of Theorem 4, deferring a complete argument to Section 6.

For simplicity, suppose that the joint distribution $F(\boldsymbol{y})$ of the $n \times 1$ random vector $\boldsymbol{Y}$ has a density $f(\boldsymbol{y})$ with bounded support $\mathscr{Y}$. Without loss of generality assume that the true coefficient equals $\beta_0 = 0$ and that $\sigma^2 = 1$. We use here the assumption of bounded support to simplify the proof; it is not used in the complete proof of Section 6.

Because $\boldsymbol{Y}$ has bounded support $\mathscr{Y}$ there is a set $B \subset \mathbb{R}^m$ such that $\left|\boldsymbol{y}'\Sigma^{-1}\boldsymbol{X}\beta\right| < 1$ for all $\beta \in B$ and $\boldsymbol{y} \in \mathscr{Y}$. For such values of β, define the auxiliary density function

$$f_\beta(\boldsymbol{y}) = f(\boldsymbol{y})\left(1 + \boldsymbol{y}'\Sigma^{-1}\boldsymbol{X}\beta\right). \tag{5}$$

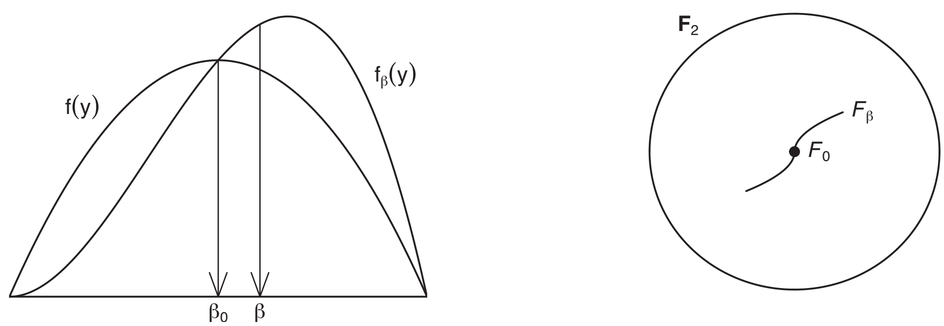
Under the assumptions, $0 \leq f_\beta(\boldsymbol{y}) \leq 2f(\boldsymbol{y})$, $f_\beta(\boldsymbol{y})$ has support $\mathscr{Y}$, and $\int_{\mathscr{Y}} f_\beta(\boldsymbol{y})d\boldsymbol{y} = 1$. To see the later, observe that $\int_{\mathscr{Y}} \boldsymbol{y}f(\boldsymbol{y})d\boldsymbol{y} = \boldsymbol{X}\beta_0 = 0$ under the normalization $\beta_0 = 0$, and thus

$$\int_{\mathscr{Y}} f_\beta(\boldsymbol{y})d\boldsymbol{y} = \int_{\mathscr{Y}} f(\boldsymbol{y})d\boldsymbol{y} + \int_{\mathscr{Y}} f(\boldsymbol{y})\boldsymbol{y}'d\boldsymbol{y}\Sigma^{-1}\boldsymbol{X}\beta = 1$$

because $\int_{\mathscr{Y}} f(\boldsymbol{y})d\boldsymbol{y} = 1$. Thus f_β is a parametric family of density functions with an associated distribution function F_β. Evaluated at β_0 we see that $f_0 = f$, which means that

F_β is a correctly-specified parametric family with true parameter value $\beta_0 = 0$.

To illustrate, take the case of a single observation with $X = 1$. Figure 1(a) displays an example density $f(y) = (3/4)(1 - y^2)$ on $[-1, 1]$ with auxiliary density $f_\beta(y) = f(y)\left(1 + y\right)$. We can see how the auxiliary density is a tilted version of the original density $f(y)$.



(a) True and Auxiliary Densities

(b) Space of Distribution Functions

Figure 1: Illustrations

Let $\mathbb{E}_\beta$ denote expectation with respect to the auxiliary distribution. Because $\int_{\mathscr{Y}} \boldsymbol{y} f(\boldsymbol{y}) d\boldsymbol{y} = 0$ and $\int_{\mathscr{Y}} \boldsymbol{y}\boldsymbol{y}' f(\boldsymbol{y}) d\boldsymbol{y} = \Sigma$, we find

$$\mathbb{E}_\beta[\boldsymbol{Y}] = \int_{\mathscr{Y}} \boldsymbol{y} f_\beta(\boldsymbol{y}) d\boldsymbol{y} = \int_{\mathscr{Y}} \boldsymbol{y} f(\boldsymbol{y}) d\boldsymbol{y} + \int_{\mathscr{Y}} \boldsymbol{y}\boldsymbol{y}' f(\boldsymbol{y}) d\boldsymbol{y} \Sigma^{-1} \boldsymbol{X}\beta = \boldsymbol{X}\beta.$$

This shows that F_β is a regression model with regression coefficient β.

In Figure 1(a), the means of the two densities are indicated by the arrows to the x-axis. In this example we can see how the auxiliary density has a larger expected value, because the density has been tilted to the right.

The parametric family F_β over $\beta \in B$ has the following properties: its expectation is $\boldsymbol{X}\beta$, its variance is finite, the true value β_0 lies in the interior of B, and the support of the distribution does not depend on β. To visualize, Figure 1(b) displays the space of finite-variance distributions $\mathbf{F}_2$ by the large circle. The dot indicates the true distribution $F = F_0$. The curved line represents the distribution family F_β. This family F_β is a sliver in the space of distributions $\mathbf{F}_2$ but includes the true distribution F.

The likelihood score of the auxiliary density function is

$$S = \frac{\partial}{\partial\beta}\log f_\beta(\boldsymbol{Y})\bigg|_{\beta=0} = \frac{\partial}{\partial\beta}\left(\log f(\boldsymbol{Y}) + \log\left(1+\boldsymbol{Y}'\Sigma^{-1}\boldsymbol{X}\beta\right)\right)\bigg|_{\beta=0} = \boldsymbol{X}'\Sigma^{-1}\boldsymbol{Y}. \tag{6}$$

Therefore the information matrix is

$$\mathscr{I} = \mathbb{E}\left[SS'\right] = \boldsymbol{X}'\Sigma^{-1}\mathbb{E}\left[\boldsymbol{Y}\boldsymbol{Y}'\right]\Sigma^{-1}\boldsymbol{X} = \boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}.$$

By assumption, $\widehat{\beta}$ is unbiased for all finite-variance distributions (the large circle in Figure 1(b)). This means that $\widehat{\beta}$ is unbiased in the subset F_β (the curve in Figure 1(b)). The Cramér-Rao lower bound states that

$$\operatorname{var}\left[\widehat{\beta}\right] \geq \mathscr{I}^{-1} = \left(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}\right)^{-1}.$$

This is the variance lower bound, completing the proof.

Some explanation may help as the argument may appear to have pulled the proverbial "rabbit out of the hat". Somehow we deduced a general variance lower bound, even though we only examined a rather artificial-looking auxiliary model. A key insight due to Stein (1956) is that the supremum of Cramér-Rao bounds over all regular parametric submodels is a lower bound on the variance of any unbiased estimator. Stein's insight focused on asymptotic variances, but the same argument applies to finite sample variances, because the Cramér-Rao bound is a finite sample result. A corollary of Stein's insight is that the Cramér-Rao bound of any single regular parametric submodel is a valid lower bound on the variance of any unbiased estimator. If this submodel is selected judiciously, its Cramér-Rao bound will equal the supremum over all submodels, and this holds when this Cramér-Rao bound equals the known finite-sample variance of a candidate efficient estimator, which in our case is the GLS estimator.

Another way of looking at this is as follows. Because $F_\beta \subset \mathbf{F}_2$, estimation over F_β cannot be harder than estimation over the full set $\mathbf{F}_2$. Thus the variance from estimation over F_β cannot be larger than estimation over $\mathbf{F}_2$. This means that Cramér-Rao bound for F_β is a lower bound for the full set $\mathbf{F}_2$.

This raises the question: How was the density (5) constructed? The trick is to construct a density which (i) includes the true density as a special case, (ii) is a regression model, and (iii) its Cramér-Rao bound equals the variance of the GLS estimator. The key is (6), which shows that the likelihood score of (5) is proportional to the score of the normal regression model with covariance matrix Σ. This was achieved by constructing (5) to be proportionate to the normal regression score.

5 Conclusion

A core question in econometric methodology is: Why do we use specific estimators? Why not others? A standard answer is *efficiency*: the estimators are best (in some sense) among all estimators (in a class) for all data distributions (in some set). The Gauss-Markov Theorem is a core efficiency result but restricts attention to linear estimators – and this is an inherently uninteresting restriction. The present paper lifts this restriction without imposing additional cost. Henceforth, least squares should be described as the "best unbiased estimator" of the regression coefficient; the "linear" modifier is unnecessary.

6 Proof of Theorem 4

We provide a proof of Theorem 4. Theorems 5 and 6 are special cases, so follow as corollaries.

Proof of Theorem 4: Our approach is to calculate the Cramér-Rao bound for a carefully crafted parametric model. This is based on an insight of Newey (1990, Appendix B) for the simpler context of a population expectation.

Without loss of generality, assume that the true coefficient equals $\beta_0 = 0$ and that $\sigma^2 = 1$. These are merely normalizations which simplify the notation.

Define the truncation function $\mathbb{R}^n \to \mathbb{R}^n$

$$\psi_c(\boldsymbol{y}) = \boldsymbol{y}\mathbb{1}\left\{\left\|\boldsymbol{y}\right\| \le c\right\} - \mathbb{E}\left[\boldsymbol{Y}\mathbb{1}\left\{\|\boldsymbol{Y}\| \le c\right\}\right]. \tag{7}$$

Notice that it satisfies $\mathbb{E}\left[\psi_c(\boldsymbol{Y})\right] = 0$,

$$\left\|\psi_c(\boldsymbol{y})\right\| \le 2c, \tag{8}$$

and

$$\mathbb{E}\left[\boldsymbol{Y}\psi_c(\boldsymbol{Y})'\right] = \mathbb{E}\left[\boldsymbol{Y}\boldsymbol{Y}'\mathbb{1}\left\{\|\boldsymbol{Y}\| \le c\right\}\right] \stackrel{\text{def}}{=} \Sigma_c.$$

As $c \to \infty$, $\Sigma_c \to \mathbb{E}\left[\boldsymbol{Y}\boldsymbol{Y}'\right] = \Sigma$. Pick c sufficiently large so that $\Sigma_c > 0$, which is feasible because $\Sigma > 0$.

Define the auxiliary joint distribution function $F_\beta(\boldsymbol{y})$ by the Radon-Nikodym derivative

$$\frac{dF_\beta\left(\boldsymbol{y}\right)}{dF\left(\boldsymbol{y}\right)} = 1 + \psi_c(\boldsymbol{y})'\Sigma_c^{-1}\boldsymbol{X}\beta$$

for parameters β in the set

$$B_c = \left\{\beta \in \mathbb{R}^m : \left\|\Sigma_c^{-1}\boldsymbol{X}\beta\right\| \le \frac{1}{4c}\right\}. \tag{9}$$

The Schwarz inequality and the bounds (8) and (9) imply that for $\beta \in B_c$ and all $\boldsymbol{y}$

$$\left|\psi_c(\boldsymbol{y})'\Sigma_c^{-1}\boldsymbol{X}\beta\right| \le \left\|\psi_c(\boldsymbol{y})\right\| \left\|\Sigma_c^{-1}\boldsymbol{X}\beta\right\| \le \frac{1}{2}.$$

This implies that F_β has the same support as F and satisfies the bounds

$$\frac{1}{2} \le \frac{dF_\beta\left(\boldsymbol{y}\right)}{dF\left(\boldsymbol{y}\right)} \le \frac{3}{2}. \tag{10}$$

We calculate that

$$\begin{aligned} \int dF_\beta\left(\boldsymbol{y}\right) &= \int dF\left(\boldsymbol{y}\right) + \int \psi_c(\boldsymbol{y})'\Sigma_c^{-1}\boldsymbol{X}\beta dF\left(\boldsymbol{y}\right) \\ &= 1 + \mathbb{E}\left[\psi_c(\boldsymbol{Y})\right]'\Sigma_c^{-1}\boldsymbol{X}\beta \\ &= 1 \end{aligned} \tag{11}$$

the last equality because $\mathbb{E}\left[\psi_c(\boldsymbol{Y})\right] = 0$. Together, these facts imply that F_β is a valid distribution function, and over $\beta \in B_c$ is a parametric family for $\boldsymbol{Y}$. Evaluated at $\beta_0 = 0$, which is in the interior of B_c, we see $F_0 = F$. This means that F_β is a correctly-specified parametric family with the true parameter value β_0.

Let $\mathbb{E}_\beta$ denote expectation under the distribution F_β. The expectation of $\boldsymbol{Y}$ in this model is

$$\begin{aligned}
\mathbb{E}_\beta[\boldsymbol{Y}] &= \int \boldsymbol{y} dF_\beta(\boldsymbol{y}) \\
&= \int \boldsymbol{y} dF(\boldsymbol{y}) + \int \boldsymbol{y}\psi_c(\boldsymbol{y})'\Sigma_c^{-1}\boldsymbol{X}\beta dF(\boldsymbol{y}) \\
&= \mathbb{E}[\boldsymbol{Y}] + \mathbb{E}\left[\boldsymbol{Y}\psi_c(\boldsymbol{Y})'\right]\Sigma_c^{-1}\boldsymbol{X}\beta \\
&= \boldsymbol{X}\beta \qquad (12)
\end{aligned}$$

because $\mathbb{E}[\boldsymbol{Y}] = 0$ and $\mathbb{E}\left[\boldsymbol{Y}\psi_c(\boldsymbol{Y})'\right] = \Sigma_c$. Thus, distribution F_β is a linear regression with regression coefficient β.

The bound (10) implies

$$\mathbb{E}_\beta\left[\|\boldsymbol{Y}\|^2\right] = \int \|\boldsymbol{y}\|^2 dF_\beta(\boldsymbol{y}) \le \frac{3}{2}\int \|\boldsymbol{y}\|^2 dF(\boldsymbol{y}) = \frac{3}{2}\mathbb{E}\left[\|\boldsymbol{Y}\|^2\right] = \frac{3}{2}\operatorname{tr}(\Sigma) < \infty.$$

This means that $F_\beta \in \mathbf{F}_2$ for all $\beta \in B_c$.

The likelihood score for F_β is

$$\begin{aligned}
S &= \frac{\partial}{\partial\beta}\log\frac{dF_\beta(\boldsymbol{Y})}{dF(\boldsymbol{Y})}\bigg|_{\beta=0} \\
&= \frac{\partial}{\partial\beta}\log\left(1+\psi_c(\boldsymbol{Y})'\Sigma_c^{-1}\boldsymbol{X}\beta\right)\bigg|_{\beta=0} \\
&= \boldsymbol{X}'\Sigma_c^{-1}\psi_c(\boldsymbol{Y}).
\end{aligned}$$

The information matrix is

$$\begin{aligned}
\mathscr{I}_c &= \mathbb{E}\left[SS'\right] \\
&= \boldsymbol{X}'\Sigma_c^{-1}\mathbb{E}\left[\psi_c(\boldsymbol{Y})\psi_c(\boldsymbol{Y})'\right]\Sigma_c^{-1}\boldsymbol{X} \\
&\le \boldsymbol{X}'\Sigma_c^{-1}\boldsymbol{X}, \qquad (13)
\end{aligned}$$

where the inequality is

$$\mathbb{E}\left[\psi_c(\boldsymbol{Y})\psi_c(\boldsymbol{Y})'\right] = \Sigma_c - \mathbb{E}\left[\boldsymbol{Y}\mathbb{1}\{\|\boldsymbol{Y}\| \le c\}\right]\mathbb{E}\left[\boldsymbol{Y}\mathbb{1}\{\|\boldsymbol{Y}\| \le c\}\right]' \le \Sigma_c.$$

By assumption, the estimator $\widehat{\beta}$ is unbiased for β for all $F \in \mathbf{F}_2$, which implies that it is unbiased for all $F \in F_\beta$. The model F_β is regular (it is correctly specified as it contains the true distribution F, the support of $\boldsymbol{Y}$ does not depend on β, and the true value $\beta_0 = 0$ lies in the interior of B_c). Thus by the Cramér-Rao Theorem (see, for example, Theorem 10.6 of Hansen (2022))

$$\operatorname{var}\left[\widehat{\beta}\right] \geq \mathscr{I}_c^{-1} \geq \left(\boldsymbol{X}'\Sigma_c^{-1}\boldsymbol{X}\right)^{-1}$$

where the second inequality is (13). Because this holds for all c, and $\Sigma_c \to \Sigma$ as $c \to \infty$,

$$\operatorname{var}\left[\widehat{\beta}\right] \geq \limsup_{c\to\infty}\left(\boldsymbol{X}'\Sigma_c^{-1}\boldsymbol{X}\right)^{-1} = \left(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}\right)^{-1}.$$

This is the variance lower bound. $\blacksquare$

References

[1] Aitken, Alexander C. (1935): "On least squares and linear combinations of observations," *Proceedings of the Royal Statistical Society*, 55, 42-48.

[2] Begun, Janet M., W. J. Hall, Wei-Min Huang, and Jon A. Wellner (1983): "Information and asymptotic efficiency in parametric-nonparametric models," *The Annals of Statistics*, 11, 432-452.

[3] Berk, Robert and Jiunn T. Hwang (1989): "Optimality of the least squares estimator," *Journal of Multivariate Analysis*, 3, 245-254.

[4] Bickel, Peter J. (1982): "On adaptive estimation," *Annals of Statistics*, 647-671.

[5] Bickel, Peter J., Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press.

[6] Chamberlain, Gary (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305-334.

[7] Doss, Hani and Jayaram Sethuraman (1989): "The price of bias reduction when there is no unbiased estimate," *The Annals of Statistics*, 17, 440-442.

[8] Gauss, Carl Friedrich (1809): *Theoria motus corporum celestium*. Hamburg: Perthese et Besser.

[9] Gauss, Carl Friedrich (1823): *Theoria Comationis Observationum Erroribus Minimis Obnoxiae*. Göttingen: Dieterich.

[10] Gnot, S., H. Knautz, G. Trenkler, and R. Zmyslony (1992): "Nonlinear unbiased estimation in linear models," *Statistics*, 23, 5-16.

[11] Hajek, Jaroslav (1972): "Local asymptotic minimax and admissibility in estimation," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 175-194.

[12] Hansen, Bruce E. (2022): *Probability and Statistics for Economists*, Princeton University Press, forthcoming.

[13] Harville, David A. (1981): "Unbiased and minimum-variance unbiased estimation of estimable functions for fixed linear models with arbitrary covariance structure," *The Annals of Statistics*, 9, 633-637.

[14] Kariya, Takeaki (1985): "A nonlinear version of the Gauss-Markov theorem," *Journal of the American Statistical Association*, 80, 476-477.

[15] Kariya, Takeaki and Hiroshi Kurata (2002): "A maximal extension of the Gauss-Markov theorem and its nonlinear version," *Journal of Multivariate Analysis*, 83, 37-55.

[16] Koenker, Roger, and Gilbert Bassett (1978): "Regression quantiles," *Econometrica*, 46, 33-50.

[17] Koopman, Reinhardt (1982): *Parameterschätzung bei a-priori-Information*, vol. 12, Vandenhoeck & Ruprecht.

[18] Laplace, Pierre Simon (1811): "Mémoire sur les integrales définies et leur application aux probabilités, et specialement à la recherche du milieu qu'il faut choisir entre les resultats des observations," *Mémoires de l'Académie des sciences de Paris*, 279-347.

[19] Legendre, Adrien Marie (1805): *Nouvelles méthodes pour la détermination des orbites des cométes*. Paris: Courcier.

[20] Lehmann, Erich L. and George Casella (1998): *Theory of Point Estimation*, Second Edition, Springer.

[21] Levit, B. Y. (1975): "On the efficiency of a class of nonparametric estimates," *Theory of Probability and its Applications*, 20, 723-740.

[22] Markov, Andreĭ Andreevich (1912): *Wahrscheinlichkeitsrechnung*. Leipzig.

[23] McDonald, James B. and Whitney K. Newey (1988): "Partially adaptive estimation of regression models via the generalized t distribution," *Econometric Theory*, 4, 428-457.

[24] Newey, Whitney K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99-135.

[25] Ritov, Ya'acov and Peter J. Bickel (1990): "Achieving information bounds in non and semiparametric models," *The Annals of Statistics*, 18, 925-938.

[26] Stein, Charles (1956): "Efficient nonparametric testing and estimation," *Berkeley Symposium on Mathematical Statistics and Probability*, 187-195.

[27] Stigler, Stephen M. (1986): *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press.

[28] van der Vaart, A.W. (1998): *Asymptotic Statistics*, Cambridge University Press.

[29] Zyskind, George, and Frank B. Martin (1969): "On best linear estimation and a general Gauss-Markov theorem in linear models with arbitrary nonnegative covariance structure," *SIAM Journal on Applied Mathematics*, 17, 1190-1202.