Gauss-Markov Theorem

Statistics · Module 8 — Multivariate Linear Models

Node 8.4

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.4 — Gauss-Markov Theorem.

The Gauss-Markov theorem establishes OLS as optimal within the class of linear unbiased estimators under second-moment assumptions. This is a finite-sample result requiring no distributional assumptions beyond the first two moments. This node depends on the linear model geometry (Node 8.3) and estimation theory (Node 3.1), and enables inference in linear models (Node 8.5).

Formal Definition

Let $Y = X\beta + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$, and $\mathrm{rank}(X) = p$.

An estimator $\tilde{\beta} = CY$ is **linear** (in Y) and **unbiased** if $CX = I_p$.

Gauss-Markov Theorem. The OLS estimator $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ is the Best Linear Unbiased Estimator (BLUE): for any linear unbiased estimator $\tilde{\beta} = CY$,

$$\mathrm{Cov}(\tilde{\beta}) - \mathrm{Cov}(\hat{\beta}) \succeq 0 \quad \text{(positive semidefinite)}.$$

Equivalently, for every $a \in \mathbb{R}^p$, $\mathrm{Var}(a^\top \tilde{\beta}) \geq \mathrm{Var}(a^\top \hat{\beta})$.

Intuition

Among all estimators that are linear functions of Y and unbiased for β, OLS has the smallest variance in every direction. The reason is geometric: OLS uses the minimum-norm linear map from Y to $\hat{\beta}$, namely the pseudoinverse $X^+ = (X^\top X)^{-1} X^\top$. Any other linear unbiased estimator adds a component D that is orthogonal to the column space of X (to preserve unbiasedness) but contributes additional variance $\sigma^2 DD^\top \succeq 0$.

Structural Role

Gauss-Markov establishes the optimality of OLS within a restricted class (linear, unbiased) under minimal assumptions (first two moments only). It is a finite-sample result — no asymptotics needed. It does not claim OLS is the best among all estimators: biased estimators (ridge, LASSO) or nonlinear estimators can have lower MSE. Under normality, OLS is additionally the MLE and achieves the Cramer-Rao bound, giving a stronger optimality. When the homoskedasticity assumption fails, GLS replaces OLS as the BLUE.

Mathematical Development

Proof of the Gauss-Markov theorem.

Let $\tilde{\beta} = CY$ be any linear unbiased estimator. Write $C = (X^\top X)^{-1}X^\top + D$ where $D = C - (X^\top X)^{-1}X^\top$.

Unbiasedness constraint: $CX = I_p$ requires $(X^\top X)^{-1}X^\top X + DX = I_p + DX = I_p$, so $DX = 0$.

Covariance computation:

$$\mathrm{Cov}(\tilde{\beta}) = \sigma^2 CC^\top = \sigma^2[(X^\top X)^{-1}X^\top + D][(X^\top X)^{-1}X^\top + D]^\top.$$

Expanding, using $DX = 0$ (so $D \cdot X(X^\top X)^{-1} = 0$, meaning the cross terms vanish):

$$\mathrm{Cov}(\tilde{\beta}) = \sigma^2(X^\top X)^{-1} + \sigma^2 DD^\top = \mathrm{Cov}(\hat{\beta}) + \sigma^2 DD^\top.$$

Since $DD^\top \succeq 0$, we have $\mathrm{Cov}(\tilde{\beta}) \succeq \mathrm{Cov}(\hat{\beta})$. $\square$

Why the cross terms vanish: $(X^\top X)^{-1}X^\top D^\top = [(X^\top X)^{-1}(DX)^\top]^\top$. Since $DX = 0$, this is zero. Geometrically, the rows of D lie in $\mathcal{C}(X)^\perp$ while the rows of $(X^\top X)^{-1}X^\top$ lie in $\mathcal{C}(X)$, making them orthogonal.

Scalar estimand version. For any linear functional $\psi = a^\top\beta$, the BLUE is $a^\top\hat{\beta}$ with variance $\sigma^2 a^\top(X^\top X)^{-1}a$. Any other linear unbiased estimator of ψ has variance at least this large.

Assumptions revisited. The Gauss-Markov theorem requires: 1. Linearity: $\mathbb{E}[Y] = X\beta$. 2. Homoskedasticity: $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$. 3. Full column rank of X.

It does not require: normality of ε, independence of observations (only uncorrelatedness and equal variance), or any knowledge of σ^2.

When assumptions fail: Generalized Least Squares.

If $\mathrm{Cov}(\varepsilon) = \sigma^2\Omega$ with known $\Omega \succ 0$ and $\Omega \neq I$, OLS is no longer BLUE. The BLUE is the GLS estimator:

$$\hat{\beta}_{\mathrm{GLS}} = (X^\top\Omega^{-1}X)^{-1}X^\top\Omega^{-1}Y.$$

Derivation. Transform the model by $\Omega^{-1/2}$: let $Y^* = \Omega^{-1/2}Y$, $X^* = \Omega^{-1/2}X$, $\varepsilon^* = \Omega^{-1/2}\varepsilon$. Then $\mathrm{Cov}(\varepsilon^*) = \sigma^2 I_n$, so OLS applied to the transformed model gives

$$\hat{\beta}_{\mathrm{GLS}} = (X^{*\top}X^*)^{-1}X^{*\top}Y^* = (X^\top\Omega^{-1}X)^{-1}X^\top\Omega^{-1}Y.$$

By Gauss-Markov applied to the transformed model, this is BLUE.

$$\mathrm{Cov}(\hat{\beta}_{\mathrm{GLS}}) = \sigma^2(X^\top\Omega^{-1}X)^{-1}.$$

Feasible GLS. When Ω is unknown, replace it by a consistent estimator $\hat{\Omega}$. The feasible GLS estimator $\hat{\beta}_{\mathrm{FGLS}} = (X^\top\hat{\Omega}^{-1}X)^{-1}X^\top\hat{\Omega}^{-1}Y$ is asymptotically equivalent to GLS but not exactly BLUE in finite samples.

Beyond BLUE: bias-variance tradeoff. Ridge regression $\hat{\beta}_\lambda = (X^\top X + \lambda I)^{-1}X^\top Y$ is biased but has MSE

$$\mathrm{MSE}(\hat{\beta}_\lambda) = \sigma^2\mathrm{tr}[(X^\top X + \lambda I)^{-2}X^\top X] + \lambda^2\beta^\top(X^\top X + \lambda I)^{-2}\beta.$$

For any $\beta \neq 0$ and $\lambda > 0$ small enough, $\mathrm{MSE}(\hat{\beta}_\lambda) < \mathrm{MSE}(\hat{\beta}) = \sigma^2\mathrm{tr}(X^\top X)^{-1}$. Gauss-Markov does not contradict this because ridge regression is biased; BLUE is optimal only within the unbiased class.

Under normality: stronger optimality. If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, then OLS is the MLE and achieves the Cramer-Rao lower bound among all unbiased estimators (not just linear ones). Under normality, BLUE = UMVUE for $a^\top \beta$.

Weighted least squares as GLS. When $\mathrm{Var}(\varepsilon_i) = \sigma^2/w_i$ with known weights $w_i > 0$, the covariance is $\mathrm{Cov}(\varepsilon) = \sigma^2 W^{-1}$ where $W = \mathrm{diag}(w_1, \ldots, w_n)$. GLS gives

$$\hat{\beta}_{\mathrm{WLS}} = (X^\top W X)^{-1} X^\top W Y,$$

which minimizes $\sum_i w_i (Y_i - x_i^\top \beta)^2$. Observations with smaller variance (larger w_i) receive more weight — they are more informative.

Efficiency loss of OLS under heteroskedasticity. When $\mathrm{Cov}(\varepsilon) = \sigma^2 \Omega \neq \sigma^2 I$, OLS remains unbiased ($\mathbb{E}[\hat{\beta}] = \beta$) but its covariance is

$$\mathrm{Cov}(\hat{\beta}_{\mathrm{OLS}}) = \sigma^2 (X^\top X)^{-1} X^\top \Omega X (X^\top X)^{-1},$$

which differs from the naive $\sigma^2 (X^\top X)^{-1}$. The difference $\mathrm{Cov}(\hat{\beta}_{\mathrm{OLS}}) - \mathrm{Cov}(\hat{\beta}_{\mathrm{GLS}}) \succeq 0$, quantifying the efficiency loss from ignoring the error structure.

Gauss-Markov for estimable functions. When X does not have full column rank, β is not identifiable, but linear functions $a^\top \beta$ that lie in $\mathcal{C}(X^\top)$ are still estimable. The Gauss-Markov theorem extends: for any estimable $a^\top \beta$, the OLS estimate (via the pseudoinverse) is BLUE.

Worked Examples

Example 1. Consider the model $Y_i = \mu + \varepsilon_i$ with $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$. Here $X = \mathbf{1}_n$, so $\hat{\mu} = \bar{Y}$ with $\mathrm{Var}(\hat{\mu}) = \sigma^2/n$. Any other linear unbiased estimator is $\tilde{\mu} = \sum c_i Y_i$ with $\sum c_i = 1$, and $\mathrm{Var}(\tilde{\mu}) = \sigma^2 \sum c_i^2 \geq \sigma^2/n$ by Cauchy-Schwarz with equality iff $c_i = 1/n$ for all i.

Example 2 (Heteroskedastic errors). Let $Y_i = \beta x_i + \varepsilon_i$ with $\mathrm{Var}(\varepsilon_i) = \sigma^2 x_i^2$ (no intercept). Here $\Omega = \mathrm{diag}(x_1^2, \ldots, x_n^2)$. The OLS estimator is $\hat{\beta}_{\mathrm{OLS}} = \sum x_i Y_i / \sum x_i^2$, but the GLS estimator is $\hat{\beta}_{\mathrm{GLS}} = \sum Y_i/x_i / \sum 1 = n^{-1} \sum Y_i/x_i$, which weights each observation inversely proportional to its variance. GLS has smaller variance than OLS in this setting.

Example 3 (Ridge vs. OLS). Consider $X^\top X = I_p$, $n = p$, $\sigma^2 = 1$, $\|\beta\|^2 = B^2$. The MSE of OLS is $\mathrm{tr}(I_p) = p$. The MSE of ridge with λ is $p/(1+\lambda)^2 + \lambda^2 B^2/(1+\lambda)^2$. Setting the derivative to zero: optimal $\lambda^* = p/B^2$, giving MSE $= pB^2/(p + B^2) < p$. Ridge strictly dominates OLS for every $B > 0$ when $p \geq 3$ (Stein's phenomenon). The Gauss-Markov theorem is not violated because ridge is biased.

Cross-Links

AI. Gauss-Markov justifies OLS in the unbiased regime; regularized methods (ridge, LASSO) trade bias for variance reduction, stepping outside the BLUE framework. The bias-variance tradeoff is the central concept in statistical learning theory.

Economics. GLS is the basis for correcting heteroskedasticity (weighted least squares) and serial correlation (Cochrane-Orcutt, Prais-Winsten) in time series econometrics. The Hausman test compares OLS and GLS to detect misspecification.

Linear Algebra. The Gauss-Markov proof uses the decomposition of the linear map $C = X^+ + D$ into the pseudoinverse plus a perturbation orthogonal to the column space. The Loewner order $\succeq$ on covariance matrices is the partial order on symmetric matrices induced by the cone of positive semidefinite matrices.

Quantum Mechanics. The BLUE concept has an analogue in quantum estimation: the locally unbiased estimator with minimum variance under the symmetric logarithmic derivative inner product achieves the quantum Cramer-Rao bound, with the same structure of projecting onto a minimum-variance subspace.

Structural Compression

Invariant. Among all linear unbiased estimators of β, OLS has the smallest covariance matrix in the Loewner order.

Mechanism. Any competing linear unbiased estimator adds a matrix D satisfying $DX = 0$; the cross terms vanish by orthogonality and the covariance increment $\sigma^2 DD^\top \succeq 0$ is nonnegative definite. The constraint $DX = 0$ forces the perturbation into the orthogonal complement of the column space.

Cross-System Echo. Gauss-Markov is the second-moment analogue of the Cramer-Rao bound: both establish lower bounds on estimator variance within a restricted class. Under normality, the two coincide. The BLUE property is a finite-sample optimality result; the Cramer-Rao bound can be either finite-sample (under regularity) or asymptotic.

Exercises

1. Verify the Gauss-Markov theorem for the intercept-only model $Y_i = \mu + \varepsilon_i$ by directly showing $\bar{Y}$ has minimum variance among all $\sum c_i Y_i$ with $\sum c_i = 1$.
2. Derive the GLS estimator by transforming the model with $\Omega^{-1/2}$ and applying OLS to the transformed model.
3. Show that if $\mathrm{Cov}(\varepsilon) = \sigma^2\Omega$ with $\Omega \neq I$, OLS is still unbiased but no longer BLUE, and compute its (incorrect) variance and its true variance.
4. Prove that ridge regression has lower MSE than OLS for λ sufficiently small, by showing $d(\mathrm{MSE}(\hat{\beta}_\lambda))/d\lambda|_{\lambda=0} < 0$ when $\beta \neq 0$.
5. In a model with two predictors and $X^\top X = \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix}$, show that Gauss-Markov implies $\hat{\beta}_1 = \bar{X}_1$ and $\hat{\beta}_2 = \bar{X}_2$ are individually optimal, and explain why orthogonality of predictors simplifies the result.
6. Construct an example where a biased estimator has strictly lower MSE than the BLUE for every value of β in a bounded parameter space.
7. Argue, structurally, why Gauss-Markov is both powerful and limited: powerful because it requires only second-moment assumptions, limited because the class of linear unbiased estimators excludes the most effective modern methods (regularization, nonlinear methods).

Statistics · Module 8 — Multivariate Linear Models · Node 8.4
Concepts: variance-and-covariance, linear-operators, orthogonality
Prerequisites: 8.3, 3.1
Enables: 8.5

hbar.university — own.your.cognition