A Modern Gauss-Markov Theorem

Bruce E. Hansen

Winter School 2022

Delhi School of Economics

OLS is BLUE

- Ordinary Least Squares (OLS) is the most popular estimator in applied econometrics.
  - The sample mean is a special case.
- In statistics and econometrics classes, it is typically taught that sample means and OLS are BLUE:
  - Best Linear Unbiased Estimator
  - This is foundational.
- The goals of today's talk:
  - A better understanding of BLUE.
  - Replace BLUE with BUE (Best Unbiased Estimator).

Gauss-Markov (BLUE) Theorems

- **BLUE**: The sample mean is the Best (minimum variance) Unbiased Linear Estimator of the population mean in the i.i.d. sampling model.
- **Gauss-Markov**: OLS is the Best Unbiased Linear Estimator of the regression coefficients in the homoskedastic linear regression model.
- Foundational efficiency result for econometric methodology.
  - Taught to students in introductory through advanced courses.
- The Gauss-Markov Theorem has not changed for 200 years.

Let's start by reviewing some of the history of least squares estimation and the Gauss-Markov Theorem

Adrien Marie Legendre (1805)

- Problem: How do we fit coefficients when there are more equations than coefficients?
- Example: Multiple observations on the location of heavenly bodies.
- Legendre's solution: Minimize the sum of squared equation errors.
- First-order condition for minimization: k linear equations with k unknown coefficients.
- The system of equations can be solved by "ordinary" methods.
- Hence the name "ordinary" least squares.
- The method was quickly adopted due to its conceptual and computational simplicity.

Carl Friedrich Gauss (1809)

- Annoyed with Legendre (1805).
  - Gauss claimed to have discovered least squares before Legendre (in unpublished uncirculated work).
- In his 1809 paper, Gauss makes several contributions:
  - He introduces a probabilistic foundation for the problem, derives the method of maximum likelihood, derives the normal distribution, and derives least squares as the maximum likelihood estimator.
- Gauss also studies the problem of multiple equations with a smaller number of unknows.
- Gauss's first innovation is to treat the equation errors as random variables. This provides a probabilistic foundation.
- Gauss introduces the method of maximum likelihood: Find the parameter values which make the observed data "most likely".
- Gauss derives the error distribution for which the MLE equals OLS. Backwards reasoning!
- We now call this distribution the "normal" or "Gaussian".

Pierre Simon Laplace (1811)

- Central limit theorem.
- Produces the normal distribution by averaging in large samples.
- Provides an alternative motivation for the assumption of normal errors.
- Examined asymptotic variance of OLS estimator.
  - Showed that OLS has the smallest asymptotic variance.
  - Asymptotic Efficiency

Gauss-Markov Theorem

- Gauss (1823)
  - Revisted earlier work.
  - Established the Gauss-Markov Theorem.
  - Homoskedastic Regression model
- Andreĭ Andreevich Markov (1912)
  - Textbook treatment of Gauss-Markov theorem.
  - Clarified the central role of unbiasedness.
- Alexander Aitken (1935)
  - Generalized the Gauss-Markov Theorem
  - Allowed for arbitrary covariance matrices.

Gauss-Markov Theorem

Theorem: OLS is the minimum variance unbiased linear estimator in the homoskedastic linear regression model.

- "minimum variance" – efficiency criterion
  - Informally, "minimum variance" = "best"
- "unbiased linear estimator" – class of estimators
  - "unbiased" – $\mathbb{E}\left[\widehat{\beta}_{\text{ols}}\right] = \beta$ for all distributions
  - "linear" – Linear in $\mathbf{Y}$
- "homoskedastic linear regression" – class of data distributions
- Goal: Can we remove **linear estimator** from the Theorem?
  - Is least squares (and sample mean) the minimum variance estimator among all unbiased estimators, including nonlinear estimators?

Linear Regression Model

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \mathbf{e} \\ \mathbb{E}\left[\mathbf{e}\right] &= 0 \\ \mathrm{var}\left[\mathbf{e}\right] &= \mathbf{\Sigma} \end{aligned}$$

- Notation: $\mathbf{Y}$ and $\mathbf{e}$ are $n \times 1$ vectors and $\mathbf{X}$ an is $n \times k$ matrix.
- We treat $\mathbf{X}$ as fixed.
- Error covariance matrix is $\mathbf{\Sigma}$.
  - We may treat $\mathbf{\Sigma}$ as unconstrained, diagonal, or i.i.d.
  - Unconstrained is for the case of general unknown covariance structure
  - Diagonal $\mathbf{\Sigma}$ arises under independent (or uncorrelated) errors
  - i.i.d. errors (homoskedasticity) imply $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_n$.

Least Squares

- The standard estimator of β is least squares.

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

- Bias

$$\mathbb{E}\left[\widehat{\beta}_{\text{ols}}\right] = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbb{E}\left[\mathbf{e}\right]\right) = \beta$$

  - $\widehat{\beta}_{\text{ols}}$ is unbiased

- Variance

$$\text{var}\left[\widehat{\beta}_{\text{ols}}\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{\Sigma}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

  - Under homoskedasticity, $\text{var}\left[\widehat{\beta}_{\text{ols}}\right] = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$

- Efficiency:
  - Is there an alternative unbiased estimator with reduced variance?

Gauss-Markov Theorem

- A "**linear estimator**" of β takes the form $\widehat{\beta} = \mathbf{A}(\mathbf{X})\mathbf{Y}$.
  - "Linear estimator" means "linear in $\mathbf{Y}$".

- **Gauss-Markov**. If $\widehat{\beta}$ is a linear estimator, and unbiased, then

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

  under homoskedasticity.

- In words:
  - No unbiased linear estimator has a finite-sample covariance matrix smaller than least squares.
  - OLS is the minimum variance **linear** unbiased estimator under homoskedasticity.

Comments on Gauss-Markov

- Beautiful in its simplicity:
  - Only assumptions on distribution are first two moments.
  - Only assumptions on estimator are linearity and unbiasedness.
  - Simple proof: easy to teach.
- Unsatisfactory Feature:
  - Restriction to **linear** estimators is unnatural.
  - No reason to exclude nonlinear estimators.
  - Double use of linear – **linear** estimators in **linear** regression – may be confusing.

Proof that Sample Mean is BLUE under Homoskedasticity

- Notation:
  - Y_i is an individual observation (i.i.d.)
  - $\mathbf{Y}$ is the $n \times 1$ vector of observations.
- Moments

$$\begin{aligned} \mu &= \mathbb{E}[Y_i] \\ \sigma^2 &= \operatorname{var}[Y_i] \\ \operatorname{var}[\mathbf{Y}] &= \sigma^2 \mathbf{I}_n \end{aligned}$$

- Sample mean

$$\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} a_i Y_i$$

where $a_i = n^{-1}$.

Proof, continued

- Linear estimators

$$\widehat{\mu} = \sum_{i=1}^{n} a_i Y_i = \mathbf{a}'\mathbf{Y}$$

  with $\mathbf{a}$ unrestricted

- Unbiased linear estimators
  - $\mathbb{E}[\widehat{\mu}] = \sum_{i=1}^{n} a_i \mathbb{E}[Y_i] = \sum_{i=1}^{n} a_i \mu = \mathbf{a}'\mathbf{1}_n \mu$ where $\mathbf{1}_n$ is an $n \times 1$ vector of 1's
  - $\mathbb{E}[\widehat{\mu}] = \mu$ if and only if $\mathbf{a}'\mathbf{1}_n = 1$
  - All unbiased linear estimators equal $\widehat{\mu} = \mathbf{a}'\mathbf{Y}$ with $\mathbf{a}$ satisfying $\mathbf{a}'\mathbf{1}_n = 1$.

Proof, continued

- For any linear estimator under homoskedasticity,

$$\operatorname{var}[\widehat{\mu}] = \operatorname{var}\left[\mathbf{a}'\mathbf{Y}\right] = \mathbf{a}' \operatorname{var}[\mathbf{Y}]\, \mathbf{a} = \sigma^2 \mathbf{a}'\mathbf{a}$$

- The linear estimator with the smallest variance is the one with smallest $\mathbf{a}'\mathbf{a}$
- Question: Which vector $\mathbf{a}$ satisfying $\mathbf{a}'\mathbf{1}_n = 1$ minimizes $\mathbf{a}'\mathbf{a}$?
  - Minimization of a quadratic subject to a linear constraint
  - Lagrangian: $\min \frac{1}{2}\mathbf{a}'\mathbf{a} - \lambda(\mathbf{a}'\mathbf{1}_n - 1)$
  - F.O.C.: $0 = \mathbf{a} - \lambda \mathbf{1}_n$
  - The solution satisfying $\mathbf{a}'\mathbf{1}_n = 1$ is $\mathbf{a} = n^{-1}\mathbf{1}_n$
- Conclusion: Best unbiased linear estimator is the sample mean

Difficulties with the BLUE Theorem

- The restriction to *linear* estimators is unnatural
- For example, the sample median is a reasonable estimator, but it is not *linear*.
- Unbiasedness is also a very special property
- Are there unbiased nonlinear estimators?

Unbiased Nonlinear Estimators

- Focus on estimation of the mean
- Assume Y_i are uncorrelated, with mean μ, variances σ_i^2
- Sample mean $\overline{Y}$ is unbiased for μ
- Nonlinear estimator
  - $\widehat{\mu} = \overline{Y} + (Y_1 - Y_2)(Y_3 - Y_4)$
  - $\mathbb{E}[\widehat{\mu}] = \mathbb{E}\left[\overline{Y}\right] + \mathbb{E}[(Y_1 - Y_2)(Y_3 - Y_4)]$
    $= \mu + \operatorname{cov}[Y_1, Y_3] + \operatorname{cov}[Y_2, Y_4] - \operatorname{cov}[Y_1, Y_4] - \operatorname{cov}[Y_2, Y_3]$
    $= \mu$
  - $\widehat{\mu}$ is unbiased
  - $\widehat{\mu}$ is a nonlinear (linear-quadratic) function of $\mathbf{Y}$
- Nonlinear unbiased estimators exist!

Unbiased Nonlinear Estimators

- Regression model
- Koopman (1982) and Gnot, Knautz, Trenkler, Zmyslony (1992) studied nonlinear unbiased estimators.
- They found that any unbiased estimator can be written as a linear-quadratic function of $\mathbf{Y}$.
- Our nonlinear example is a linear-quadratic function of $\mathbf{Y}$.
- $\widehat{\mu} = \overline{Y} + (Y_1 - Y_2)(Y_3 - Y_4)$

What is the Best Unbiased Estimator?

- What is the best unbiased estimator, allowing for nonlinear estimators?
- A possible answer can be found in the Cramér-Rao Theorem

Cramér-Rao Theorem

Theorem: Let $\mathbf{Y}$ have joint density $f(\mathbf{y}, \theta)$ [plus mild regularity]. Then, if an estimator $\widehat{\theta}$ is unbiased for θ,

$$\operatorname{var}[\widehat{\theta}] \geq \mathcal{I}_n^{-1}$$

where

$$\mathcal{I}_n = \operatorname{var}\left[\frac{\partial}{\partial \theta} \log f\left(\mathbf{Y}, \theta\right)\right]$$

is the Fisher Information.

Cramér-Rao: Interpretation

$$\text{var}[\widehat{\theta}] \geq \mathcal{I}_\theta^{-1}$$

- The quantity $\mathcal{I}_\theta^{-1}$ is a *variance lower bound.*
- No unbiased estimator can have a lower variance than $\mathcal{I}_\theta^{-1}$.
- This is also the asymptotic variance of the MLE, so it is effectively the lowest feasible variance among unbiased estimators.

Cramér-Rao: Proof

- Assume θ is scalar.
- Since $f(\mathbf{y}, \theta)$ is a density for all θ,

$$1 = \int f(\mathbf{y}, \theta) d\mathbf{y}$$

- Take the derivative of the two sides with respect to θ,

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial \theta} f(\mathbf{y}, \theta) d\mathbf{y} \\ &= \int \frac{\partial}{\partial \theta} \log f(\mathbf{y}, \theta) f(\mathbf{y}, \theta) d\mathbf{y} \\ &= \mathbb{E}\left[\frac{\partial}{\partial \theta} \log f(\mathbf{Y}, \theta)\right] \end{aligned}$$

- Thus $\mathbb{E}\left[\frac{\partial}{\partial \theta} \log f(\mathbf{Y}, \theta)\right] = 0$

Cramér-Rao: Proof, continued

- Let $\widehat{\theta} = \widehat{\theta}(\mathbf{Y})$ be unbiased. Then

$$\theta = \mathbb{E}[\widehat{\theta}] = \int \widehat{\theta}(\mathbf{y}) f(\mathbf{y}, \theta) d\mathbf{y}$$

- Take the derivative of the two sides with respect to θ,

$$\begin{aligned} 1 &= \int \widehat{\theta}(\mathbf{y}) \frac{\partial}{\partial \theta} f(\mathbf{y}, \theta) d\mathbf{y} \\ &= \int \widehat{\theta}(\mathbf{y}) \frac{\partial}{\partial \theta} \log f(\mathbf{y}, \theta) f(\mathbf{y}, \theta) d\mathbf{y} \\ &= \mathbb{E}\left[\widehat{\theta} \frac{\partial}{\partial \theta} \log f(\mathbf{Y}, \theta)\right] \\ &= \mathbb{E}\left[(\widehat{\theta} - \theta) \frac{\partial}{\partial \theta} \log f(\mathbf{Y}, \theta)\right] \end{aligned}$$

Cramér-Rao: Proof, continued

- We have shown $1 = \mathbb{E}\left[(\widehat{\theta} - \theta)\dfrac{\partial}{\partial\theta}\log f(\mathbf{Y}, \theta)\right]$
- Squaring, then using the Cauchy-Schwarz inequality,

$$\begin{aligned} 1 &= \left(\mathbb{E}\left[(\widehat{\theta} - \theta)\frac{\partial}{\partial\theta}\log f(\mathbf{Y}, \theta)\right]\right)^2 \\ &\leq \operatorname{var}[\widehat{\theta}]\operatorname{var}\left[\frac{\partial}{\partial\theta}\log f(\mathbf{Y}, \theta)\right] \\ &= \operatorname{var}[\widehat{\theta}]\,\mathcal{I}_n \end{aligned}$$

- Rewriting,

$$\operatorname{var}[\widehat{\theta}] \geq \mathcal{I}_n^{-1}$$

- This is the Cramér-Rao inequality.

Application 1: Normal Regression

- $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$
- $\dfrac{\partial}{\partial \beta} \log f\left(\mathbf{Y}, \beta\right) = \dfrac{1}{\sigma^2}\mathbf{X}'\mathbf{e}$
- $\mathcal{I}_n = \mathrm{var}\left[\dfrac{\partial}{\partial \beta} \log f\left(\mathbf{Y}, \beta\right)\right] = \dfrac{1}{\sigma^2}\mathbf{X}'\mathbf{X}$
- $\mathcal{I}_n^{-1} = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$
- Thus, if $\widehat{\beta}$ is unbiased for β, then $\mathrm{var}[\widehat{\beta}] \geq \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

- Removes "linear estimator"
  - OLS is the minimum variance unbiased estimator under **normality**.
- But, normality is restrictive.

Application 2: Laplace Regression

- $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$
- $f(e) = \dfrac{1}{\sqrt{2}\sigma} \exp\left(-\dfrac{\sqrt{2}\,|e|}{\sigma}\right)$
- $\dfrac{\partial}{\partial \beta} \log f\left(\mathbf{Y}, \beta\right) = \dfrac{\sqrt{2}}{\sigma}\mathbf{X}' sgn(\mathbf{e})$
- $\mathcal{I}_n = \text{var}\left[\dfrac{\partial}{\partial \beta} \log f\left(\mathbf{Y}, \beta\right)\right] = \dfrac{2}{\sigma^2}\mathbf{X}'\mathbf{X}$
- $\mathcal{I}_n^{-1} = \dfrac{\sigma^2}{2}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$
- Thus, if $\widehat{\beta}$ is unbiased for β, then $\text{var}[\widehat{\beta}] \geq \dfrac{\sigma^2}{2}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$.
- A smaller lower bound than under normality!
- Implication: If error is Laplace, OLS is not efficient

Common Misunderstanding

- If the error is normal, then OLS is efficient
- If the error is Laplace, then OLS is not efficient
- Efficiency bound depends on error distribution
- This is a misunderstanding, because the error distribution is **unknown** and **unknowable**
- If you change the estimator, the unbiased property can disappear
- For example, the LAD estimator is unbiased and inconsistent if the true density is skewed

Semi-Parametric Efficiency

- Take any model $f(\mathbf{Y}, \theta)$ which is correctly specified (in the sense that the true density is a member of the model)
- Take any unbiased estimator $\widehat{\theta}$
- The Cramér-Rao bound states that

$$\operatorname{var}[\widehat{\theta}] \geq \left(\operatorname{var}\left[\frac{\partial}{\partial \theta} \log f(\mathbf{Y}, \theta) \right] \right)^{-1}$$

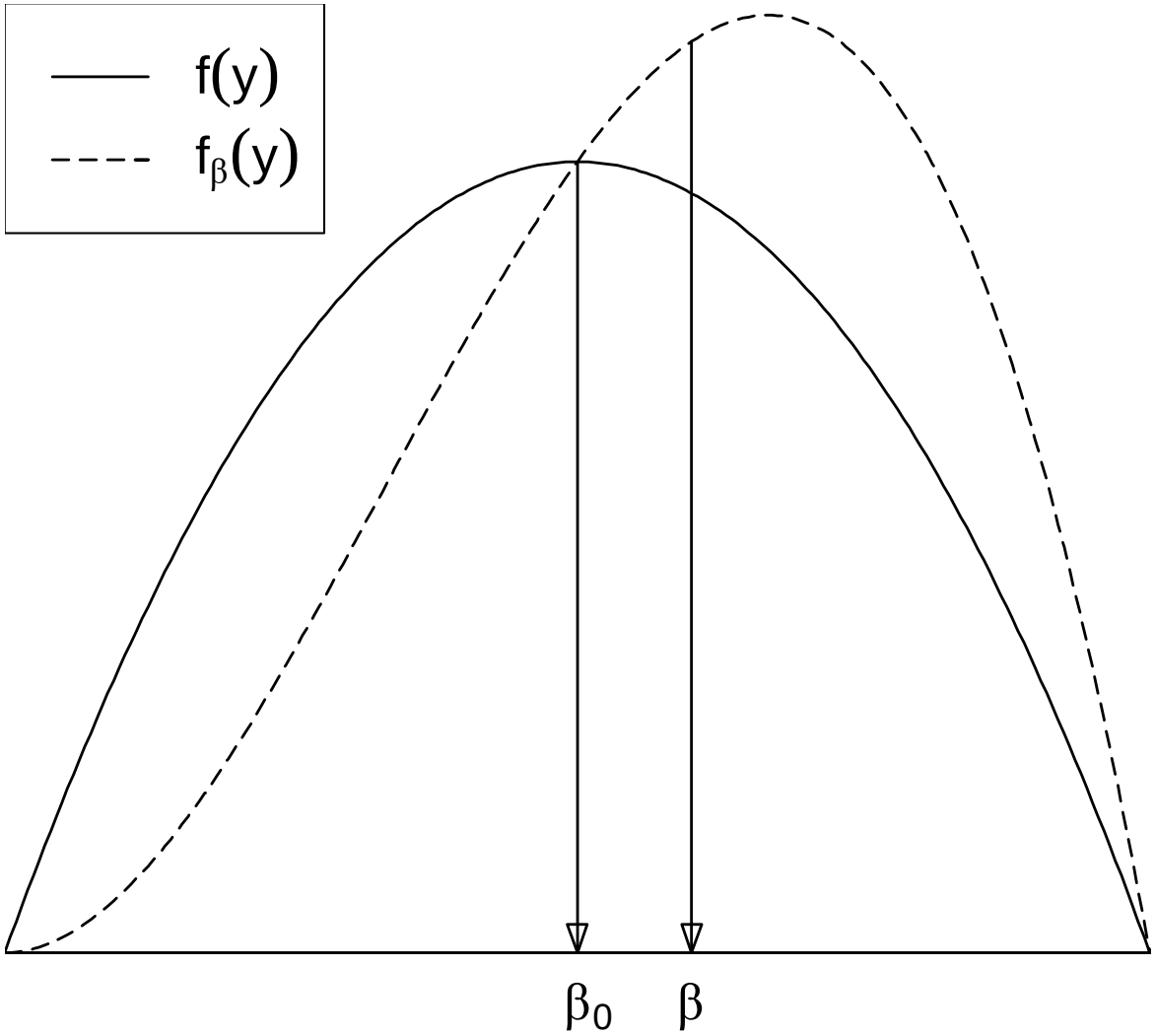
- This holds for all valid models, thus $\operatorname{var}[\widehat{\theta}]$ must be larger than all such lower bounds.
- The semi-parametric bound is found by finding the worst case – the largest lower bound.
- Short-cut: Find a model $f(\mathbf{Y}, \theta)$ with a convenient Cramér-Rao bound.

Application 3: Tilted Density

- Estimation of the mean, i.i.d. sample
  - $\mu_0 = \mathbb{E}[Y]$
  - $\sigma^2 = \text{var}[Y]$
  - Y has unknown density $f(y)$, so $\int f(y)dy = 1$
  - WLOG, assume $\mu_0 = 0$, so $\int yf(y)dy = 0$.
- Model:
  - $f(y,\mu) = f(y)\left(1 + \frac{y}{\sigma^2}\mu\right)$
  - $\int f(y,\mu)dy = \int f(y)dy + \int yf(y)dy\frac{\mu}{\sigma^2} = 1$
    - ★ $f(y,\mu)$ is a valid density, so this is a valid model
  - $\int yf(y,\mu)dy = \int yf(y)dy + \int y^2 f(y)dy\frac{\mu}{\sigma^2} = \mu$
    - ★ Thus, $f(y,\mu)$ has mean μ
    - ★ At $\mu = \mu_0 = 0$, $f(y,\mu_0) = f(y)$ is the true density
    - ★ Thus, $f(y,\mu)$ is a correctly specified model

Example of Tilted Density

- $f(y) = \frac{3}{4}\left(1 - y^2\right)$ on $[-1, 1]$
- $f(y, \mu) = f(y)(1 + y\mu) = \frac{3}{4}\left(1 - y^2\right)(1 + y\mu)$



Application 3: Tilted Density, continued

- $f(y,\mu) = f(y)\left(1+\dfrac{y}{\sigma^2}\mu\right)$
- $\dfrac{\partial}{\partial\mu}\log f(y,\mu) = \dfrac{\partial}{\partial\mu}\log f(y) + \dfrac{\partial}{\partial\mu}\log\left(1+\dfrac{y}{\sigma^2}\mu\right) = \dfrac{\frac{y}{\sigma^2}}{1+\frac{y}{\sigma^2}\mu}$
- At $\mu = \mu_0 = 0$, $\dfrac{\partial}{\partial\mu}\log f(y,\mu_0) = \dfrac{y}{\sigma^2}$
- $\mathcal{I}_n = \operatorname{var}\left[\dfrac{\partial}{\partial\mu}\log f\left(Y,\mu_0\right)\right] = n\dfrac{\sigma^2}{\sigma^4} = n\sigma^{-2}$
- Thus, if $\widehat{\mu}$ is unbiased in $f(y,\mu)$, then $\operatorname{var}[\widehat{\mu}] \geq \mathcal{I}_n^{-1} = \dfrac{\sigma^2}{n}$
- Holds if $\widehat{\mu}$ is unbiased for i.i.d. samples

Lower Bound for Estimation of the Mean

- We have shown that for i.i.d. Y_i with mean μ, variance σ^2, if $\widehat{\mu}$ is unbiased for μ, then $\operatorname{var}[\widehat{\mu}] \geq \dfrac{\sigma^2}{n}$
- The sample mean $\overline{Y}$ satisfies $\operatorname{var}[\overline{Y}] = \dfrac{\sigma^2}{n}$
- Hence $\operatorname{var}[\widehat{\mu}] \geq \operatorname{var}[\overline{Y}]$.
- $\overline{Y}$ has lowest variance among unbiased estimators

Best Unbiased Estimator

Theorem: Best Unbiased Estimator (BUE)
For i.i.d. Y_i, if $\widehat{\mu}$ is unbiased for $\mu = \mathbb{E}[Y]$, then

$$\operatorname{var}[\widehat{\mu}] \geq \operatorname{var}[\overline{Y}] = \frac{\sigma^2}{n}$$

Discussion

- We have proved that the sample mean is BUE in addition to BLUE
- The only assumption was that the observations are i.i.d.
- The proof shown here used the technical assumption that Y has a density, but this can be relaxed
- Therefore, under i.i.d. sampling, the sample mean is the best (minimum variance) unbiased estimator, without the restriction to linear estimators, and without the restriction to the normal distribution.

Common Confusion

- Other estimators can be more efficient than the sample mean, in some cases.
- For example, when the errors are Laplace, then the sample median has lower variance than the sample mean.
- The confusion is that these estimators are not unbiased.
- "Unbiased" means that the estimator has zero bias for all distributions in the model class (for any distribution with i.i.d. sampling).
- The sample median, for example, is a biased estimator of the population mean when the distribution is skewed.
- The reduction in variance obtained by specialized estimators only occurs for specialized distributions, and this comes at the cost of bias, otherwise.

Heterogeneous Variances

- Y_i are independent, with mean μ, variances σ_i^2
- Best linear unbiased estimator

$$\widetilde{\mu} = \frac{\sum_{i=1}^n \sigma_i^{-2} Y_i}{\sum_{i=1}^n \sigma_i^{-2} Y_i}$$

- It has variance

$$\mathrm{var}[\widetilde{\mu}] = \left(\sum_{i=1}^n \sigma_i^{-2} \right)^{-1}$$

- Use a similar argument as before, but with tilted densities

$$f_i(y, \mu) = f_i(y) \left(1 + \frac{y}{\sigma_i^2} \mu \right)$$

Best Unbiased Estimator with Heterogeneous Variances

Theorem: Best Unbiased Estimator (BUE)
For independent Y_i with common mean μ but heterogeneous variances σ_i^2, if $\widehat{\mu}$ is unbiased for μ, then

$$\operatorname{var}[\widehat{\mu}] \geq \left(\sum_{i=1}^{n} \sigma_i^{-2} \right)^{-1}$$

Discussion

- Context: Observations are independent but have heterogeneous variances
- The BUE is the "GLS" weighted average

$$\widetilde{\mu} = \frac{\sum_{i=1}^{n} \sigma_i^{-2} Y_i}{\sum_{i=1}^{n} \sigma_i^{-2} Y_i}$$

- This is the best unbiased estimator among all possibly nonlinear estimators, across all error distributions
- This estimator is infeasible, as the variances σ_i^2 are unknown, but establishes the ideal estimator

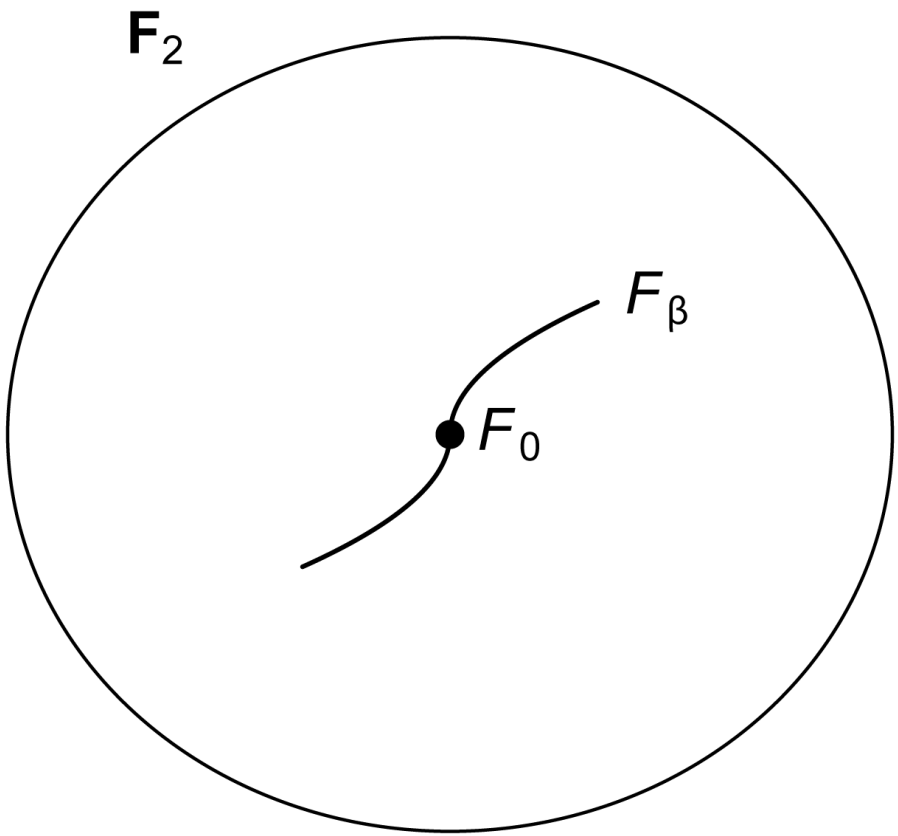
Explanation for BUE Theorem

- Stein (1956) observed 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 focused on asymptotic variances, but the same argument applies to finite sample variances.
- A corollary 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.
- This can be verified when this Cramér-Rao bound equals the known finite-sample variance of a candidate efficient estimator.
- For estimation of the mean this is the sample mean or weighted mean.

Illustration of sub-models

- $\mathbf{F}_2$ is set of all error distributions with finite variance
- F_0 is true distribution
- F_β is tilted distribution
  - $F_0 \in F_\beta$
  - $F_\beta \subset \mathbf{F}_2$

Illustration



Discussion

- The idea is that the class F_β is a valid model class
  - It is a valid distribution
  - It is a subset of the general class $\mathbf{F}_2$
  - The true distribution F_0 is a member of F_β
- By the Cramer-Rao theorem, we can find the variance lower bound for the class F_β
  - This is σ^2 / n
- If an estimator is unbiased throughout $\mathbf{F}_2$ then it is also unbiased throughout F_β
- By the Cramer-Rao theorem, this estimator cannot have a variance lower than σ^2 / n

Regression with Uncorrelated Observations

- $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$
- Errors e_i mutually uncorrelated, possibly heteroskedastic
- **Gauss-Markov (BLUE)**: If variances are homoskedastic, and $\widehat{\beta}$ is linear and unbiased

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

- **Modern Gauss-Markov (BUE)**: If observations are mutually independent, and $\widehat{\beta}$ is unbiased, then under homoskedasticity,

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

Discussion

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

- The classical Gauss-Markov theorem restricted attention to linear estimators
- The modern Gauss-Markov theorem removes the restriction to linear estimators
- The BUE theorem is not a strict improvement, as it requires that the observations are independent, not just uncorrelated, and requires that $\widehat{\beta}$ is unbiased under heteroskedasticity

Do Nonlinear Unbiased Estimators Exist for Regression?

- Take the independent sampling framework
- For some i, let $\widehat{\beta}_{-i}$ be the leave-one-out estimator (estimation without observation i
- For some $j \neq i$, set

$$\widetilde{\beta} = \widehat{\beta}_{\text{ols}} + Y_i \left(Y_j - X_j' \widehat{\beta}_{-i} \right)$$

- This estimator $\widetilde{\beta}$ is a nonlinear (quadratic) function of $\mathbf{Y}$.

Nonlinear Estimator, continued

$$\widetilde{\beta} = \widehat{\beta}_{\text{ols}} + Y_i \left(Y_j - X_j' \widehat{\beta}_{-i} \right)$$

- It is unbiased
  - $\widehat{\beta}_{-i}$ is unbiased: $\mathbb{E}\left[\widehat{\beta}_{-i}\right] = \beta$
  - $\mathbb{E}\left[Y_j - X_j' \widehat{\beta}_{-i}\right] = \mathbb{E}\left[X_j' \beta + e_j - X_j' \widehat{\beta}_{-i}\right] = 0$
  - $Y_j - X_j' \widehat{\beta}_{-i}$ is not a function of (Y_i, X_i) so they are independent.

$$\begin{aligned} \mathbb{E}\left[\widetilde{\beta}\right] &= \mathbb{E}\left[\widehat{\beta}_{\text{ols}}\right] + \mathbb{E}\left[Y_i \left(Y_j - X_j' \widehat{\beta}_{-i} \right)\right] \\ &= \beta + \mathbb{E}\left[Y_i\right] \mathbb{E}\left[Y_j - X_j' \widehat{\beta}_{-i}\right] \\ &= \beta \end{aligned}$$

- Thus nonlinear unbiased estimators exist, under independent sampling

Regression with Heteroskedatic Variances

- $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$
- Errors e_i mutually uncorrelated, possibly heteroskedastic
- $\mathrm{var}[\mathbf{e}] = \mathbf{D}$, diagonal matrix
- GLS estimator (if $\mathbf{D}$ known)
  - $\widehat{\beta}_{\mathrm{gls}} = \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{Y}\right)$
- Properties
  - $\widehat{\beta}_{\mathrm{gls}}$ is unbiased
  - $\mathrm{var}\left[\widehat{\beta}_{\mathrm{gls}}\right] = \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}$

BLUE and BUE

- **Aitken's Theorem (BLUE)**: If $\widehat{\beta}$ is linear and unbiased, then

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

- **Modern Gauss-Markov (BUE)**: If observations are mutually independent, and $\widehat{\beta}$ is unbiased, then

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

Discussion

- BUE removes the requirement that the estimator is linear.
- BUE requires that the observations are independent.

General Covariance Matrix

- $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$
- $\text{var}[\mathbf{e}] = \mathbf{\Sigma}$, unstructured
- GLS estimator
  - $\widehat{\beta}_{\text{gls}} = \left(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{Y}\right)$
- Properties
  - $\widehat{\beta}_{\text{gls}}$ is unbiased
  - $\text{var}\left[\widehat{\beta}_{\text{gls}}\right] = \left(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-1}$

BLUE and BUE

- **Aitken's Theorem:** If $\widehat{\beta}$ is linear and unbiased, then

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

- **Modern Gauss-Markov (BUE)**: If $\widehat{\beta}$ is unbiased, then

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

- BUE Theorem is a strict generalization

However: Unbiased Estimators are Linear

- $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$
- $\mathrm{var}[\mathbf{e}] = \mathbf{\Sigma}$, unstructured
- Stephen Portnoy "Linearity of Unbiased Linear Model Estimators" (*American Statistician*, 2022)
- Benedikt Pötscher and David Preinerstorfer "A Modern Gauss-Markov Theorem? Really?" (working paper)
- These papers prove that if an estimator $\widehat{\beta}$ is unbiased for unstructured $\mathbf{\Sigma}$, then $\widehat{\beta}$ must be linear in $\mathbf{Y}$.

Unbiased Estimators are Linear

- Portnoy and Pötscher-Preinerstorfer, unstructured $\mathbf{\Sigma}$
- They provide is an alternative proof of Aitken's Theorem:
  - Classical: If $\widehat{\beta}$ is linear and unbiased for all $\mathbf{\Sigma}$, then $\operatorname{var}\left[\widehat{\beta}\right] \geq \left(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-1}$
  - New: If $\widehat{\beta}$ is unbiased for all $\mathbf{\Sigma}$, then it is linear
  - Combined: If $\widehat{\beta}$ is unbiased for all $\mathbf{\Sigma}$, then $\operatorname{var}\left[\widehat{\beta}\right] \geq \left(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-1}$
- This result is confined to the case of unstructure $\mathbf{\Sigma}$.

Other Comments and Extensions

- H D Vinod "Material Facts Obscured in Hansen's Modern Gauss-Markov Theorem" (March 13, 2022)
  - Points out that most interesting estimators are nonlinear, and biased, so focus on unbiased estimators is misguided.
- Gib Basset "The Modern Gauss Markov Theorem for the Median" (March 10, 2022)
  - The sample median is a median-unbiased estimator of the population median. In what sense is it the best estimator?
  - Lemma: In the class of median-unbiased estimators of the population median, the one with smallest dispersion is the sample median.
- Gib Basset "The Gauss Markov Theorem for the Least Absolute Error Estimator" (March 10, 2022)
  - Median regression model $Y = X'\beta + e$ with $Med[e \mid X] = 0$
  - Claim: LAD estimator has smallest dispersion among median-unbiased estimators of β

Summary

- We teach the Gauss-Markov Theorem to students.
  - "**Least squares is the best linear unbiased estimator**."
- We can simplify the statement:
  - "**Least squares is the best unbiased estimator**."