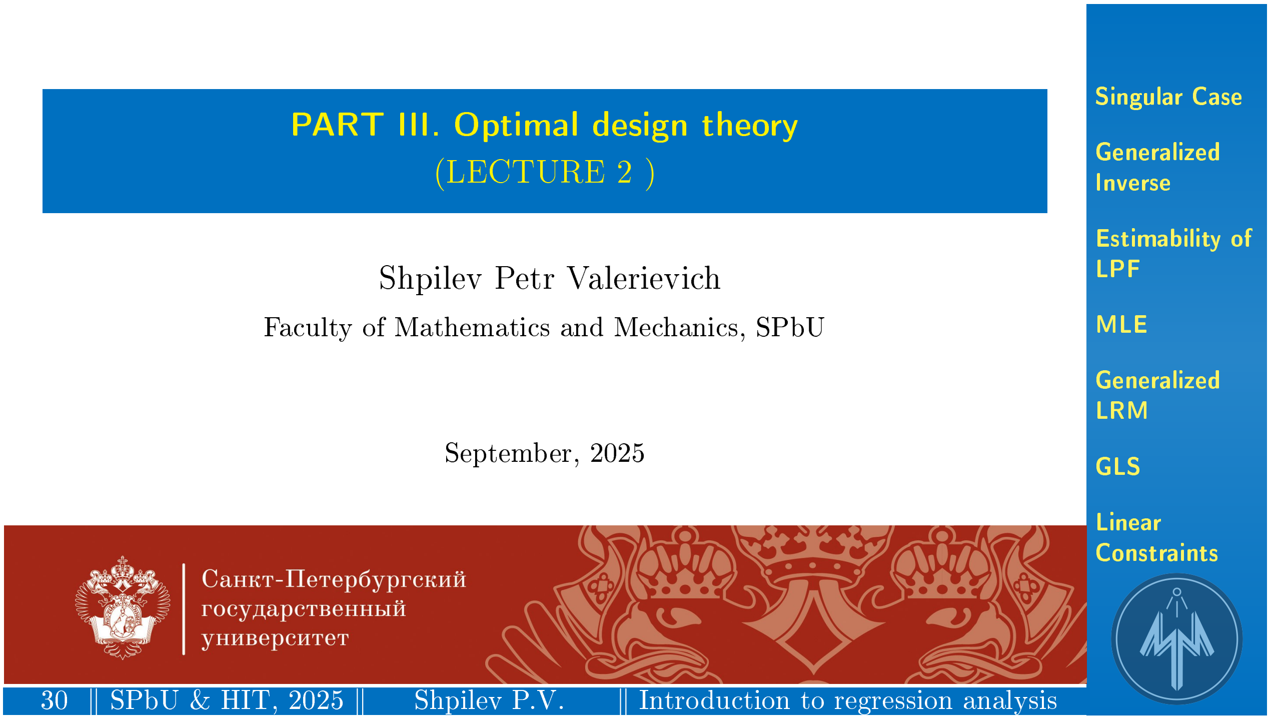


PART III. Optimal design theory
(LECTURE 2)

Shpilev Petr Valerievich
Faculty of Mathematics and Mechanics, SPbU

September, 2025

Санкт-Петербургский
государственный
университет

Comments

In this lecture, we extend the theory of regression estimation to more general and challenging settings. We begin with the ordinary least squares (OLS) estimator in the singular case, introducing the concept of the generalized inverse, its definition, existence, and key properties, with special attention to the Moore–Penrose inverse. This framework allows us to address estimation when the design matrix is singular and to revisit the estimability of linear parametric functions, illustrated through a detailed example from weighing experiments.

We then turn to maximum likelihood estimation (MLE) in linear regression, proving its equivalence with OLS under normal assumptions and highlighting classical results such as the IID normal sample as a special case. The lecture continues with the generalized linear regression model and the derivation of the generalized least squares (GLS) estimator, which accommodates correlated and heteroscedastic errors.

Finally, we consider regression under additional structural information, including prior knowledge of parameters, linear constraints, and estimation within restricted parameter sets. This leads to the study of minimax estimators under quadratic constraints, their special cases, and their optimality in the class of unbiased estimators. The lecture concludes with a Bayesian perspective, introducing estimators based on a known prior distribution.

OLS Estimation in the Singular Case

Singular Case
Generalized Inverse
Estimability of LPF
MLE
Generalized LRM
GLS
Linear Constraints

Motivation

In many applications, including analysis of variance (ANOVA), the matrix $X^T X$ is singular. Therefore, generalizing OLS estimation to this case is essential.

- So far, we have assumed that the matrix $X^T X$ is non-singular.
- Now we address the case where the matrix $X^T X$ is singular, i.e., $\mathrm{rank}(X^T X) < m$.
- In this case, the normal equation
$$X^T X\theta = X^T Y$$
still has solutions, but not a unique one.
- Any vector $\widehat{\theta}$ that satisfies the normal equation minimizes the residual sum of squares $\|Y - X\theta\|^2$, and is called an OLS estimator.

Key Point

In the singular case, the OLS estimator exists but is not unique: the solution set of the normal equations is infinite. These solutions can be represented using generalized inverse matrices.

Comments

Let us now transition to the general case where the matrix $X^T X$ is singular. This situation arises frequently in applied settings, especially in analysis of variance, where the design matrix includes linearly dependent columns. Such linear dependencies make the matrix $X^T X$ non-invertible.

Despite this complication, we still define the ordinary least squares estimator as any vector $\widehat{\theta}$ that minimizes the squared norm of $Y - X\theta$. According to a previously proven Lemma 1, this vector $\widehat{\theta}$ satisfies the normal equation: $X^T X\widehat{\theta} = X^T Y$. However, because the matrix on the left is singular, the normal equation no longer has a unique solution. Instead, it has infinitely many solutions.

Each of these solutions provides the same minimum value of the residual sum of squares, but the vector $\widehat{\theta}$ itself is not uniquely defined. This distinguishes the singular case from the non-singular one, where the least squares estimator is given explicitly by the inverse of $X^T X$ times $X^T Y$.

This transition marks an important shift in focus. We no longer seek a unique estimator, but instead investigate which functions of θ remain estimable in the presence of singularity. This perspective will guide the development of generalized estimation theory for linear models.

Generalized vs. Moore–Penrose Inverse

Singular Case

Generalized Inverse

Estimability of LPF

MLE

Generalized LRM

GLS

Linear Constraints

Definition

Let $A \in \mathbb{R}^{n\times m}$. A matrix $A^{-} \in \mathbb{R}^{m\times n}$ is called a generalized inverse of A if, for every vector $y \in \mathbb{R}^{n}$ such that the system $Ax = y$ is consistent, the vector $x = A^{-}y$ is a solution.

Remarks

- Generalized inverse matrices are not unique.
- Any matrix admits at least one generalized inverse.
- Generalized inverses allow representation of all solutions of consistent linear systems.

Moore–Penrose Pseudoinverse

A matrix $A^{+} \in \mathbb{R}^{m\times n}$ is called the Moore–Penrose pseudoinverse of A if it satisfies all four Penrose conditions:

$$\begin{aligned} &(1)\quad AA^{+}A = A, && (3)\quad (AA^{+})^{\top} = AA^{+}, \\ &(2)\quad A^{+}AA^{+} = A^{+}, && (4)\quad (A^{+}A)^{\top} = A^{+}A. \end{aligned}$$

The pseudoinverse always exists and is unique.

Comments

We now introduce the notion of a generalized inverse. Let A be a real matrix of size $n \times m$. A matrix A^{-} of size $m \times n$ is called a generalized inverse of A if it maps every vector y in $\mathbb{R}^{n}$, for which the equation $Ax = y$ is consistent, to a solution $x = A^{-}y$. This concept is used to represent all possible solutions of linear systems in the case when A is not invertible or not square. It is important to note that generalized inverses are not uniquely defined — a matrix may admit infinitely many generalized inverses.

Among generalized inverses, a particularly important role is played by the Moore–Penrose pseudoinverse, denoted by A^{+}. This matrix is uniquely defined for any real matrix A and satisfies four algebraic conditions, known as the Penrose equations. These conditions ensure symmetry and minimality properties that make the pseudoinverse especially valuable in linear estimation theory. In particular, the pseudoinverse yields the solution of minimal Euclidean norm to a consistent linear system.

Every Moore–Penrose pseudoinverse is a generalized inverse, but not every generalized inverse is a pseudoinverse. This distinction is essential in the theory of linear models, where various estimation criteria may select different generalized inverses.

Generalized Inverse: Definition and Property

Lemma 6: Condition for Generalized Inverse

For a matrix B to be a *generalized inverse* of matrix A, it is necessary and sufficient that B satisfies the equality:

$$\mathrm{ABA} = \mathrm{A}.$$

Proof:

Sufficiency:

- Assume a generalized inverse $\mathrm{B} = \mathrm{A}^-$ exists for matrix A.
- Consider the i-th column of matrix A, denoted as $\mathrm{a_i}$.
- The system $\mathrm{Ax} = \mathrm{a_i}$ is obviously consistent, as a solution exists (e.g., $\mathrm{x} = \mathrm{e_i}$).
- By definition, the vector $\mathrm{x} = \mathrm{A^- a_i}$ is a particular solution to this system.
- Therefore, substituting this solution into the system gives:

$$\mathrm{AA^- a_i} = \mathrm{a_i}, \quad \text{for all i}.$$

- This equality, holding for every column $\mathrm{a_i}$ of A, directly implies that $\mathrm{AA^-A} = \mathrm{A}$.

Comments

This lemma provides a necessary and sufficient condition for a matrix to be a generalized inverse. Specifically, a matrix B is a generalized inverse of a matrix A if and only if the product $\mathrm{ABA} = \mathrm{A}$.

To prove sufficiency, assume that a generalized inverse exists. Denote this inverse by A^-. We examine each column of the matrix A — let $\mathrm{a_i}$ be the i-th column. The system $\mathrm{Ax} = \mathrm{a_i}$ is clearly consistent; for example, it is solved by the i-th standard basis vector.

According to the definition, the product $\mathrm{A^- a_i}$ must also be a solution. Substituting this into the original system yields: $\mathrm{AA^- a_i} = \mathrm{a_i}$. Since this holds for every column of A, we conclude that $\mathrm{AA^-A} = \mathrm{A}$.

This condition captures the essential role of the generalized inverse: it reproduces any vector in the image of A via this triple product. It also demonstrates that the generalized inverse does not require the matrix to be square or invertible.

Generalized Inverse: Proof (continued)

Necessary$(\mathrm{ABA} = \mathrm{A} \implies \exists\ \mathrm{B} = \mathrm{A}^{-})$:

- Now, suppose the system $\mathbf{Ax} = \mathbf{y}$ is consistent, and matrix B satisfies the equality $\mathbf{ABA} = \mathbf{A}$.
- Multiply both sides of this equality by x from the right:

$$\mathbf{AB}\underbrace{\mathbf{Ax}}_{\mathbf{y}} = \underbrace{\mathbf{Ax}}_{\mathbf{y}} \quad \Rightarrow \quad \mathbf{ABy} = \mathbf{y}.$$

- This result means that the vector $\mathbf{x} = \mathbf{By}$ is a solution to the consistent system $\mathbf{Ax} = \mathbf{y}$.
- By definition, if for every consistent system $\mathrm{Ax} = \mathrm{y}$, By is a solution, then matrix B is a *generalized inverse* for A.
- Thus, $\mathrm{B} = \mathrm{A}^{-}$.

□

Remark

The condition $\mathrm{ABA} = \mathrm{A}$ is fundamental. Unlike the regular inverse, generalized inverses exist for any matrix, regardless of its shape or rank. This ensures that if a linear system $\mathrm{Ax} = \mathrm{y}$ has a solution, By will provide one.

Comments

Now we prove the necessity. Suppose that the matrix B satisfies the condition $\mathrm{ABA} = \mathrm{A}$. Consider a consistent system $\mathrm{Ax} = \mathrm{y}$. That means there exists a vector x for which $\mathrm{Ax} = \mathrm{y}$.

Multiplying both sides of the identity $\mathrm{ABA} = \mathrm{A}$ on the right by this vector x, we obtain $\mathrm{ABAx} = \mathrm{Ax}$. Since $\mathrm{Ax} = \mathrm{y}$, this implies that $\mathrm{ABy} = \mathrm{y}$. Hence, the vector $\mathrm{x} = \mathrm{By}$ is a solution of the system $\mathrm{Ax} = \mathrm{y}$.

Therefore, the matrix B provides a solution to every consistent system. By definition, this means that B is a generalized inverse of A.

In conclusion, the condition $\mathrm{ABA} = \mathrm{A}$ is both necessary and sufficient for B to be a generalized inverse.

The remark at the end emphasizes the generality of this concept. Unlike the ordinary inverse, which only exists for square and nonsingular matrices, a generalized inverse exists for any matrix, regardless of its dimensions or rank. It guarantees that for every consistent system $\mathrm{Ax} = \mathrm{y}$, the product By gives a solution.

Existence of Generalized Inverse

Lemma 7: Existence of Generalized Inverse

For any matrix A, there exists a generalized inverse A^-.

Proof:

Part 1: Symmetric Square Matrices

- Consider the case where A is a *symmetric square matrix*.
- By the *spectral decomposition theorem*, A can be expressed as:

$$\mathrm{A} = \mathrm{P}\Lambda\mathrm{P}^{\mathrm{T}}, \quad \text{where} \quad \mathrm{P}^{\mathrm{T}}\mathrm{P} = \mathrm{P}\mathrm{P}^{\mathrm{T}} = \mathrm{I}.$$

- Here, Λ is a diagonal matrix containing the eigenvalues of A.
- Without loss of generality, assume Λ has the form:

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix},$$

  where Λ_1 is a diagonal matrix with **nonzero** entries.
- Now, we define a candidate for the generalized inverse:

$$\mathrm{A}^- = \mathrm{P}\Lambda^-\mathrm{P}^{\mathrm{T}}, \quad \text{where} \quad \Lambda^- = \begin{pmatrix} \Lambda_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Comments

This slide addresses the existence of a generalized inverse for any matrix. We begin by considering the special case when the matrix A is square and symmetric. According to the spectral decomposition theorem, any symmetric matrix A can be represented as the product of three matrices: $\mathrm{A} = \mathrm{P}\Lambda\mathrm{P}^{\mathrm{T}}$, where P is an orthogonal matrix such that the product of its transpose and itself equals the identity matrix, that is, $\mathrm{P}^{\mathrm{T}}\mathrm{P} = \mathrm{I}$, and Λ is a diagonal matrix containing the eigenvalues of A.

Without loss of generality, we assume that the matrix Λ is partitioned into a block form where the top-left block, denoted by Λ_1, contains all the nonzero eigenvalues on its diagonal, while the bottom-right block consists of zeros.

We then construct a matrix Λ^- by taking the reciprocal of each nonzero diagonal entry of Λ_1 and leaving all other entries as zero. Using this, we define a candidate for the generalized inverse of A as follows: $\mathrm{A}^- = \mathrm{P}\Lambda^-\mathrm{P}^{\mathrm{T}}$.

This construction uses the same eigenbasis as A, and the diagonal inversion is only applied to the nonzero spectrum, which ensures stability in the presence of zero eigenvalues.

Existence of Generalized Inverse (continued)

Verification for Symmetric Case:

- Let's verify that the defined A^- satisfies the condition $AA^-A = A$:

$$AA^-A = P\Lambda P^T P\Lambda^- P^T P\Lambda P^T = P\Lambda\Lambda^-\Lambda P^T = P\Lambda P^T = A.$$

- Therefore, by Lemma 6, A^- is indeed a generalized inverse.

Part 2: General Case (Arbitrary Matrices)

- For any arbitrary matrix A, there exists a representation (e.g., Singular Value Decomposition) in the form **A=BΛC**.
- Here, **B** and **C** are *non-singular* matrices, and Λ is a diagonal matrix.
- Similar to the symmetric case, we can write Λ and define Λ^- as:

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \Lambda^- = \begin{pmatrix} \Lambda_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

- Now, we define the generalized inverse for A as: $A^- = C^{-1}\Lambda^- B^{-1}$.
- Let's verify that for this A^-, the condition $AA^-A = A$ holds:

$$AA^-A = B\Lambda CC^{-1}\Lambda^- B^{-1}B\Lambda C = B\Lambda\Lambda^-\Lambda C = B\Lambda C = A.$$

- Therefore, by Lemma 6, this A^- is also a generalized inverse.

The Lemma is proved. □

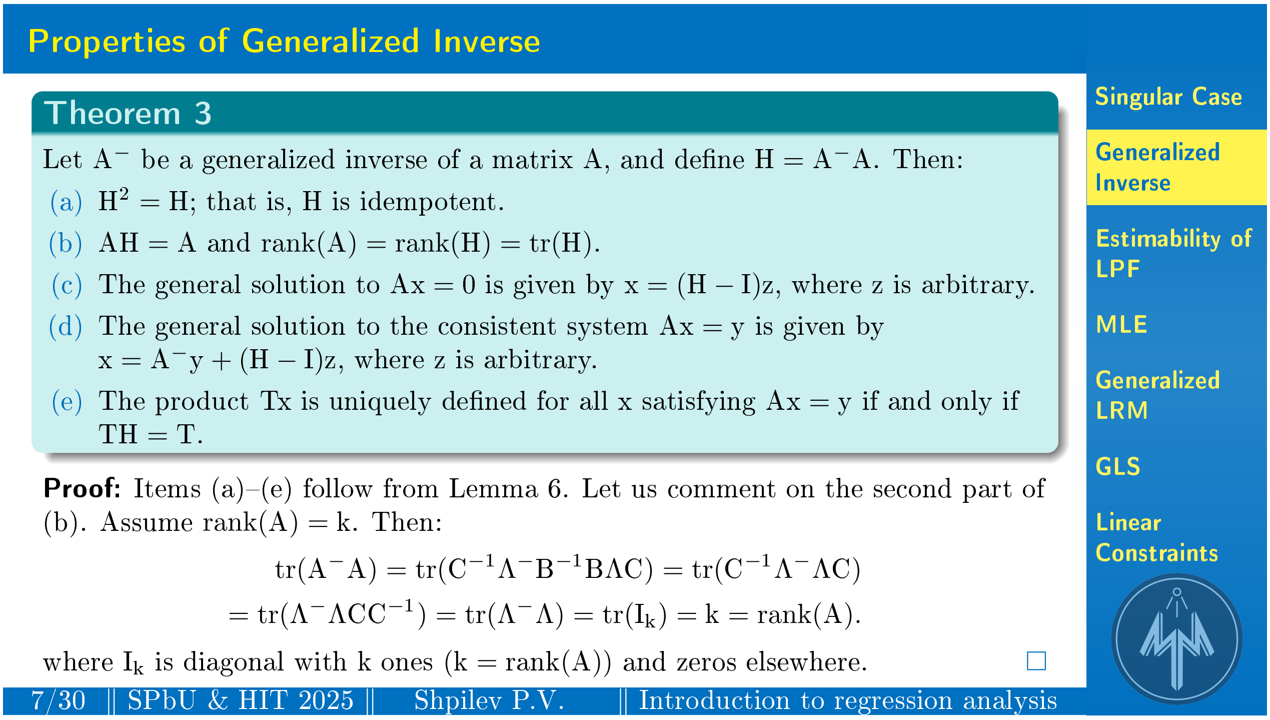
Comments

We now verify that the matrix constructed in the symmetric case truly satisfies the condition for being a generalized inverse. We compute the product AA^-A, which becomes $P\Lambda P^T$ multiplied by $P\Lambda^- P^T$ multiplied by $P\Lambda P^T$. Using the fact that the transpose of P times P equals the identity matrix, this expression simplifies to $P\Lambda\Lambda^-\Lambda P^T$. Since $\Lambda\Lambda^-\Lambda = \Lambda$, the final result is $P\Lambda P^T$, which is equal to A. Thus, the required condition is satisfied, and the constructed matrix is indeed a generalized inverse.

To extend this result to arbitrary matrices, we use a general factorization, such as the singular value decomposition. In this approach, any matrix A can be written as the product of three matrices: $A = B\Lambda C$, where B and C are nonsingular matrices and Λ is a diagonal matrix.

Again, we assume that Λ has a block structure where the top-left block Λ_1 contains nonzero diagonal entries, and the rest are zero. Then, we define Λ^- by inverting the nonzero entries of Λ_1 and leaving the rest as zero. Using this, we define the generalized inverse of A as $A^- = C^{-1}\Lambda^- B^{-1}$.

Finally, we verify the condition $AA^-A = A$. Multiplying out, we get $B\Lambda C$ multiplied by $C^{-1}\Lambda^- B^{-1}$ multiplied by $B\Lambda C$. This simplifies to $B\Lambda\Lambda^-\Lambda C$, which again equals $B\Lambda C$, and this is precisely A. Thus, this construction works for any matrix.



Properties of Generalized Inverse

Theorem 3

Let A^- be a generalized inverse of a matrix A, and define $H = A^-A$. Then:

(a) $H^2 = H$; that is, H is idempotent.

(b) $AH = A$ and $\text{rank}(A) = \text{rank}(H) = \text{tr}(H)$.

(c) The general solution to $Ax = 0$ is given by $x = (H - I)z$, where z is arbitrary.

(d) The general solution to the consistent system $Ax = y$ is given by $x = A^-y + (H - I)z$, where z is arbitrary.

(e) The product Tx is uniquely defined for all x satisfying $Ax = y$ if and only if $TH = T$.

Proof: Items (a)–(e) follow from Lemma 6. Let us comment on the second part of (b). Assume $\text{rank}(A) = k$. Then:

$$\begin{aligned}\text{tr}(A^-A) &= \text{tr}(C^{-1}\Lambda^-B^{-1}B\Lambda C) = \text{tr}(C^{-1}\Lambda^-\Lambda C)\\ &= \text{tr}(\Lambda^-\Lambda CC^{-1}) = \text{tr}(\Lambda^-\Lambda) = \text{tr}(I_k) = k = \text{rank}(A).\end{aligned}$$

where I_k is diagonal with k ones ($k = \text{rank}(A)$) and zeros elsewhere. □

Comments

This slide presents a fundamental theorem about the structure of generalized inverses. Let A be an arbitrary matrix, and let A^- be any of its generalized inverses. We define the matrix H as A^-A.

The theorem consists of five key statements.

First, item (a): the matrix H is idempotent. That means, if we multiply H by itself, we obtain H again.

Item (b): multiplying A by H on the right gives A. Moreover, the rank of A equals the rank of H, and this also equals the trace of H. The trace, being the sum of the diagonal elements, gives a numerical measure of the rank in this context.

Item (c): the general solution of the homogeneous system $Ax = 0$ is given by the formula $x = (H - I)z$, where z is an arbitrary vector.

Item (d): the general solution to the consistent system $Ax = y$ is given by $A^-y + (H - I)z$, where z is an arbitrary vector.

Finally, item (e): the linear expression Tx takes a unique value for all x satisfying $Ax = y$, if and only if $TH = T$.

The proof follows directly from the characterization of generalized inverses. Let us only comment on the trace identity in item (b). Suppose the rank of A is equal to k. Then, using a canonical factorization of A, we compute the trace of A^-A. This is equal to the trace of a matrix product involving the inverse of C, a diagonal matrix Λ^-, and the matrices B and C. After simplification using cyclicity of trace and orthogonality properties, we obtain the trace of $\Lambda^-\Lambda$, which is simply the trace of the identity matrix of order k. Therefore, the result equals k, which is the rank of A.

Singular Design Matrix: Challenges & Solutions

Background and Motivation

- In the proof, we used the well-known formula: $\mathrm{tr}(\mathrm{AB}) = \mathrm{tr}(\mathrm{BA})$ (provided the matrices are conformable).
- If design matrix X is ***singular*** ($\mathrm{rank}(\mathrm{X}) < \mathrm{m}$):
  - Unbiased estimates for all parameters θ are ***impossible***.
  - However, ***certain parametric functions*** $\tau = \mathrm{T}\theta$ ***can still be estimated***.

Why Generalized Inverses?

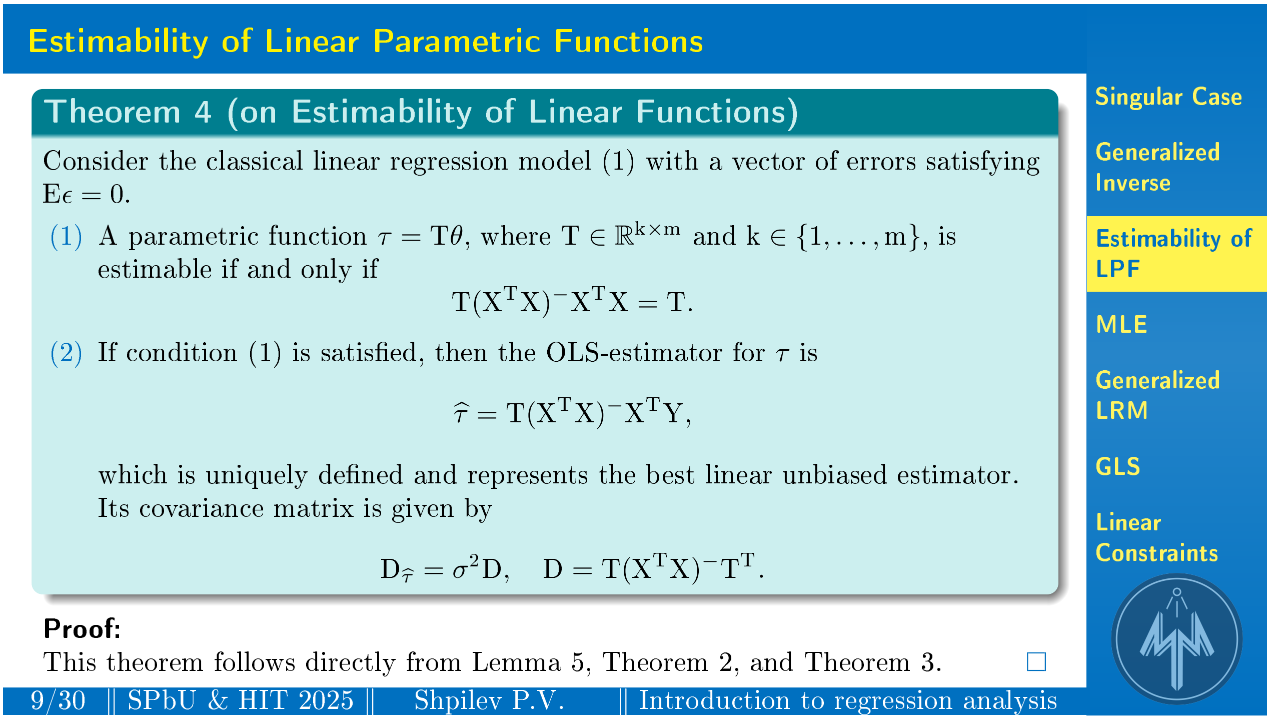
- The Moore-Penrose pseudoinverse might not preserve properties like unbiasedness.
- ***Generalized inverses*** allow constructing estimators that meet minimal conditions for ***unbiasedness*** and ***consistency***.
- They provide the necessary flexibility while maintaining $\mathrm{ABA} = \mathrm{A}$, which is crucial in deriving estimator formulas.

Comments

In the proof, we used the well-known equality: the trace of a matrix product AB equals the trace of BA, provided the matrices are conformable.

In regression settings, when the design matrix X has rank strictly less than the number of columns m, the full parameter vector θ is no longer estimable in an unbiased way. However, some linear combinations of θ, written as $\tau = \mathrm{T}\theta$, may still be estimable. Determining which combinations are estimable becomes a central concern when working with singular models.

At this point, let us clarify why generalized inverses, rather than the Moore–Penrose pseudoinverse, are often used in theoretical estimation. The Moore–Penrose pseudoinverse provides the unique solution of minimal Euclidean norm, which is suitable for numerical computations. However, this solution is not always unbiased and may not minimize variance under statistical constraints. Generalized inverses offer the flexibility to construct estimators that meet specific conditions of unbiasedness or optimality. This is why in linear model theory, generalized inverses are preferred over pseudoinverses when analyzing estimability.



Estimability of Linear Parametric Functions

Theorem 4 (on Estimability of Linear Functions)

Consider the classical linear regression model (1) with a vector of errors satisfying $\mathrm{E}\epsilon = 0$.

(1) A parametric function $\tau = \mathrm{T}\theta$, where $\mathrm{T} \in \mathbb{R}^{\mathrm{k}\times\mathrm{m}}$ and $\mathrm{k} \in \{1, \dots, \mathrm{m}\}$, is estimable if and only if

$$\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{X} = \mathrm{T}.$$

(2) If condition (1) is satisfied, then the OLS-estimator for τ is

$$\widehat{\tau} = \mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{Y},$$

which is uniquely defined and represents the best linear unbiased estimator. Its covariance matrix is given by

$$\mathrm{D}_{\widehat{\tau}} = \sigma^2\mathrm{D}, \quad \mathrm{D} = \mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{T}^{\mathrm{T}}.$$

Proof:
This theorem follows directly from Lemma 5, Theorem 2, and Theorem 3. □

Comments

This theorem addresses the problem of estimating a linear function of the parameter vector in the classical regression model. Assume that the vector of errors has expected value equal to zero. We consider a parametric function τ defined as $\mathrm{T}\theta$, where T is a matrix of size $\mathrm{k} \times \mathrm{m}$.

According to the theorem, this function τ is estimable — that is, there exists an unbiased linear estimator — if and only if the matrix identity $\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{X} = \mathrm{T}$ holds. In words: τ is estimable if and only if the product $\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{X}$ equals T.

If this condition is satisfied, then the function τ admits a unique best linear unbiased estimator. It is given by the formula: $\widehat{\tau} = \mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$.

Moreover, the covariance matrix of this estimator is equal to $\sigma^2\mathrm{D}$, where D is defined as $\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{T}^{\mathrm{T}}$.

This result generalizes the classical Gauss–Markov theorem to the case where the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is singular. In this setting, not all components of θ may be estimable, but linear functions satisfying the stated matrix identity remain estimable with minimal variance.

Example: Regression from Weighing Experiments (1/4)

Example: Model setup

Consider a regression model corresponding to the weighing of three objects using a two-pan balance. The result of each weighing is the difference in weight between the left and right pans plus random noise.

- Let $\mathrm{Y} \in \mathbb{R}^{\mathrm{N}}$ be the vector of observed differences.
- Let $\theta = (\theta_1, \theta_2, \theta_3)^{\mathrm{T}}$ be the unknown weights.
- The design matrix X has entries

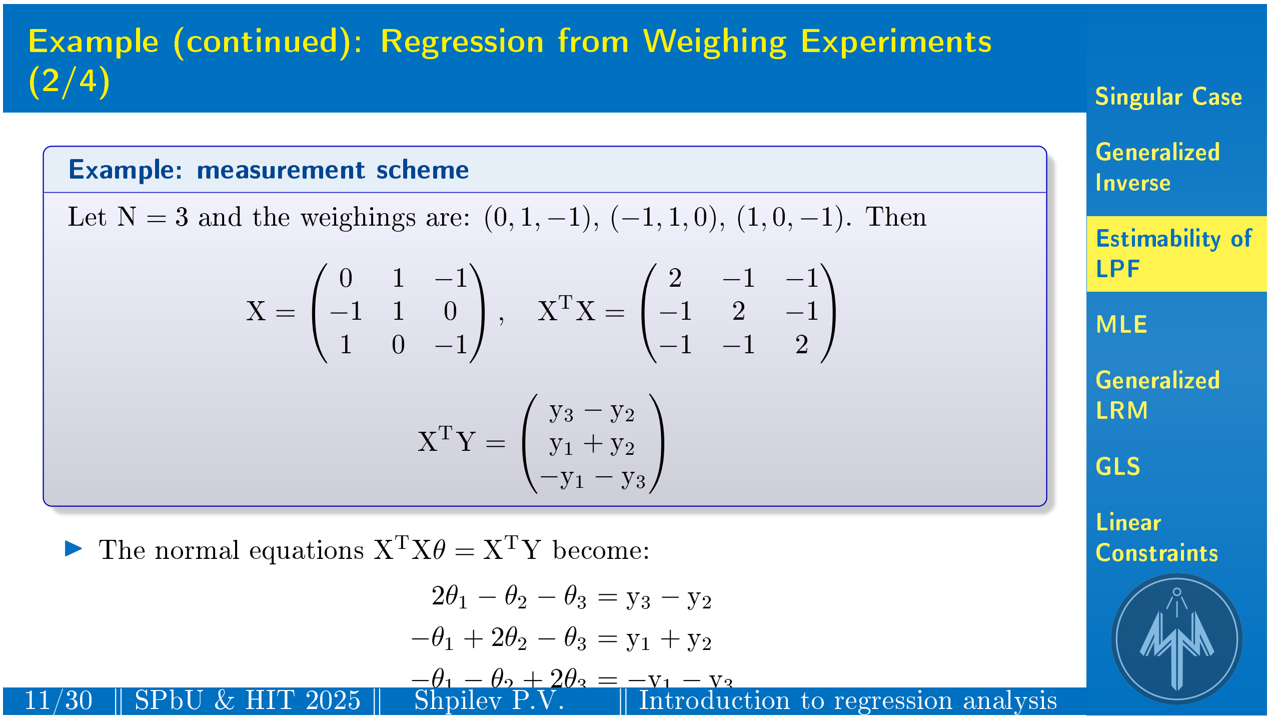
$$\mathrm{x_{ij}} = \begin{cases} 1, & \text{object j on the left pan in weighing i,} \\ -1, & \text{object j on the right pan in weighing i,} \\ 0, & \text{otherwise.} \end{cases}$$

- Regression model: $\mathrm{Y} = \mathrm{X}\theta + \epsilon$, where $\mathrm{E}\epsilon = 0$.

Comments

This example considers a classical weighing problem with three objects and a two-pan scale. The outcome of each weighing is the difference in total mass between the left and the right pan, plus a random error. We assume that the error vector satisfies the standard regression assumptions: first, the expected value of the error is zero; second, the errors have constant variance; and third, they are uncorrelated. The regression equation can be written in matrix form as $\mathrm{Y} = \mathrm{X}\theta + \epsilon$, where Y is the vector of observed differences, X is the design matrix, θ is the vector of unknown object weights, and ϵ is the vector of random errors.

Each row of the design matrix X represents a weighing configuration. For element $\mathrm{x_{ij}}$, we assign the value one if the j-th object is on the left pan during the i-th weighing, the value minus one if it is on the right pan, and zero if it is excluded from the weighing.



Example (continued): Regression from Weighing Experiments (2/4)

Example: measurement scheme

Let $N = 3$ and the weighings are: $(0, 1, -1)$, $(-1, 1, 0)$, $(1, 0, -1)$. Then

$$X = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix}, \quad X^T X = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

$$X^T Y = \begin{pmatrix} y_3 - y_2 \\ y_1 + y_2 \\ -y_1 - y_3 \end{pmatrix}$$

- The normal equations $X^T X \theta = X^T Y$ become:

$$\begin{aligned} 2\theta_1 - \theta_2 - \theta_3 &= y_3 - y_2 \\ -\theta_1 + 2\theta_2 - \theta_3 &= y_1 + y_2 \\ -\theta_1 - \theta_2 + 2\theta_3 &= -y_1 - y_3 \end{aligned}$$

Comments

In our case, there are three weighings. The first configuration places object two on the left and object three on the right, corresponding to the row $(0, 1, -1)$. The second weighing puts object one on the right and object two on the left, giving $(-1, 1, 0)$. The third weighing puts object one on the right and object three on the left, resulting in the row $(-1, 0, 1)$.

Given the design matrix X from the weighing scheme, we compute the matrix product $X^T X$, which yields a symmetric three-by-three matrix with diagonal elements equal to two and off-diagonal elements equal to minus one. We also compute $X^T Y$, which results in a three-dimensional column vector: the first element is $y_3 - y_2$, the second is $y_1 + y_2$, and the third is $-y_1 - y_3$.

Using these, we write the normal equations of the regression model: $2\theta_1 - \theta_2 - \theta_3 = y_3 - y_2$; $-\theta_1 + 2\theta_2 - \theta_3 = y_1 + y_2$; and $-\theta_1 - \theta_2 + 2\theta_3 = -y_1 - y_3$.

Example: Regression from Weighing Experiments (3/4)

Example (continued): General solution of the normal system

Solving the normal equations yields:

$$\widehat{\theta}_1 = \frac{2y_3 - y_2 + y_1}{3} + \theta_3, \quad \widehat{\theta}_2 = \frac{y_2 + 2y_1 + y_3}{3} + \theta_3$$

- The parameter θ_3 remains arbitrary.
- The system is not of full rank, so $X^T X$ is singular.
- Use of a generalized inverse is necessary for estimation.

Why generalized inverse?

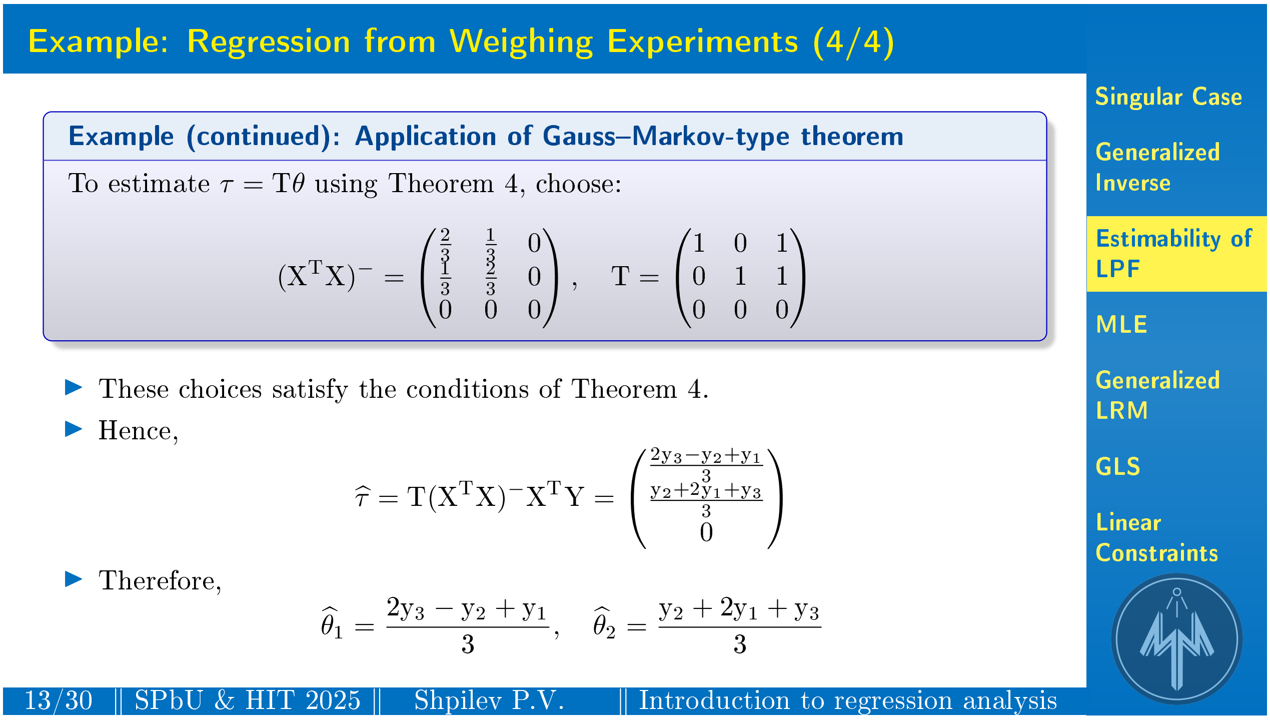
In underdetermined models, the Moore–Penrose pseudoinverse may not yield estimators for identifiable parameter functions. Generalized inverses allow for flexible estimation of identifiable combinations such as $T\theta$.

Comments

Solving the system of normal equations gives the following expressions for the least squares estimators: $\widehat{\theta}_1 = \frac{2y_3 - y_2 + y_1}{3} + \theta_3$; and $\widehat{\theta}_2 = \frac{y_2 + 2y_1 + y_3}{3} + \theta_3$. These formulas show that both estimators depend on θ_3, which remains arbitrary. This arbitrariness reflects the fact that the design matrix X does not have full rank, so the matrix $X^T X$ is singular. Therefore, the system has infinitely many solutions.

In such cases, we estimate only those linear functions of the parameter vector θ that are estimable. This means we focus on combinations like $\theta_1 + \theta_3$, or $\theta_2 + \theta_3$, for which unbiased linear estimators exist. To compute these, we need to choose a generalized inverse of the matrix $X^T X$.

One might ask why we do not simply use the Moore–Penrose pseudoinverse. The reason is practical: the Moore–Penrose inverse yields one specific solution — the one with minimal Euclidean norm — but it is not required in our setup. Any generalized inverse satisfying the standard conditions can be used to compute the best linear unbiased estimator of an estimable function. In our example, it is convenient to choose a generalized inverse that yields simple expressions for the estimators of the combinations we care about. The Moore–Penrose inverse, though perfectly valid, would result in more complicated formulas without improving the statistical properties of the estimators.



Example: Regression from Weighing Experiments (4/4)

Example (continued): Application of Gauss–Markov-type theorem

To estimate $\tau = \mathrm{T}\theta$ using Theorem 4, choose:

$$(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-} = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathrm{T} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

- These choices satisfy the conditions of Theorem 4.
- Hence,

$$\widehat{\tau} = \mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{Y} = \begin{pmatrix} \frac{2y_3 - y_2 + y_1}{3} \\ \frac{y_2 + 2y_1 + y_3}{3} \\ 0 \end{pmatrix}$$

- Therefore,

$$\widehat{\theta}_1 = \frac{2y_3 - y_2 + y_1}{3}, \quad \widehat{\theta}_2 = \frac{y_2 + 2y_1 + y_3}{3}$$

Comments

To resolve the identifiability issue in our underdetermined regression model, we now apply Theorem 4, which provides the best linear unbiased estimator for a linear transformation of the parameter vector. In our case, we wish to estimate the vector τ, defined as $\mathrm{T}\theta$, where T is a three-by-three matrix designed to extract identifiable combinations. Specifically, the first row of T corresponds to $\theta_1 + \theta_3$, and the second row to $\theta_2 + \theta_3$. The third row is all zeros, since θ_3 is not estimable on its own.

To apply the theorem, we must select a generalized inverse of the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$. We choose a matrix whose first two rows contain two-thirds and one-third in symmetric positions, and whose third row is entirely zero. This choice ensures that θ_3 remains arbitrary, while θ_1 and θ_2 are estimable through appropriate linear combinations.

Substituting these matrices into the formula from Theorem 4, we compute the best linear unbiased estimator for τ as $\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$. The result is a three-dimensional vector, where the first element is $\frac{2y_3 - y_2 + y_1}{3}$, the second element is $\frac{y_2 + 2y_1 + y_3}{3}$, and the third element is zero.

Hence, the corresponding estimates for θ_1 and θ_2 are uniquely determined, while θ_3 remains unidentifiable. This shows how the use of a generalized inverse, together with the transformation matrix T, allows us to extract the estimable parts of the parameter vector, even when the model is not of full rank.

Maximum Likelihood Estimation in Linear Regression

The Likelihood Function in Linear Regression

- Let's consider the classic linear regression model (1):

$$\mathrm{Y} = \mathrm{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

- In this model, the observed data vector Y is a sample drawn from a distribution.
- The ***probability density function*** (or probability mass function for discrete data) of Y, parameterized by the unknown parameters $\boldsymbol{\theta}$, is denoted as $\mathrm{L}(\mathrm{Y}, \boldsymbol{\theta})$.
- This function, $\mathrm{L}(\mathrm{Y}, \boldsymbol{\theta})$, is known in mathematical statistics as the ***Likelihood Function***.

Definition: Maximum Likelihood Estimator (MLE)

The value $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ that ***maximizes*** the likelihood function $\mathrm{L}(\mathrm{Y}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is called the *Maximum Likelihood Estimator*.

Singular Case
Generalized Inverse
Estimability of LPF
MLE
Generalized LRM
GLS
Linear Constraints

Comments

On this slide, we begin our study of the maximum likelihood method – a powerful tool for parameter estimation in statistical models.

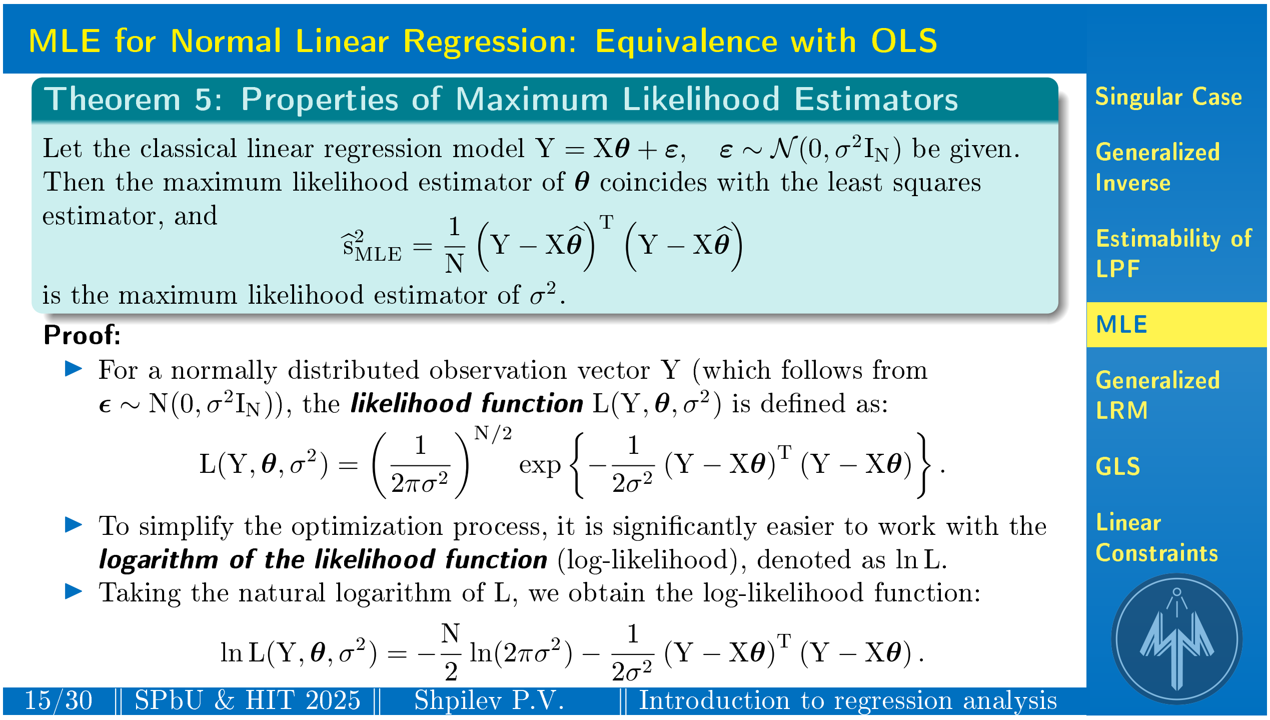
We already know that the quality of regression analysis largely depends on the accuracy of parameter estimation. The least squares method (LSM) is a common and reliable approach, but it's not the only one. When we know the error distribution law (for example, normal distribution), we can use this additional information to find the most probable values of unknown parameters (θ) given the observed data.

The function $\mathrm{L}(\mathrm{Y}, \theta)$, written as $\mathrm{L}(\mathrm{Y}, \theta)$, denotes the likelihood function. This is the probability density function of the observed data Y, treated as a function of the unknown parameters θ. In essence, the likelihood function measures how "likely" a given value of θ is, in light of the data we observed.

The maximum likelihood estimator, denoted $\widehat{\theta}_{\mathrm{MLE}}$, is defined as the value of θ that maximizes this likelihood function. That is, it is the value of the parameter vector under which the observed data are most probable.

However, the maximum likelihood framework is more general. It allows for flexible modeling assumptions, such as non-normal error distributions, heteroskedasticity, or even discrete outcomes. On the other hand, maximum likelihood estimation requires stronger distributional assumptions and may be computationally more complex than least squares, especially in high-dimensional or non-linear models.

Thus, while both methods often yield similar estimates in linear models, maximum likelihood provides a richer and more principled framework when full probabilistic modeling is desired.



MLE for Normal Linear Regression: Equivalence with OLS

Theorem 5: Properties of Maximum Likelihood Estimators

Let the classical linear regression model $Y = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_N)$ be given. Then the maximum likelihood estimator of $\boldsymbol{\theta}$ coincides with the least squares estimator, and

$$\widehat{s}^2_{MLE} = \frac{1}{N}\left(Y - X\widehat{\boldsymbol{\theta}}\right)^T \left(Y - X\widehat{\boldsymbol{\theta}}\right)$$

is the maximum likelihood estimator of σ^2.

Proof:

- For a normally distributed observation vector Y (which follows from $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_N)$), the ***likelihood function*** $L(Y, \boldsymbol{\theta}, \sigma^2)$ is defined as:

$$L(Y, \boldsymbol{\theta}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left\{-\frac{1}{2\sigma^2}(Y - X\boldsymbol{\theta})^T (Y - X\boldsymbol{\theta})\right\}.$$

- To simplify the optimization process, it is significantly easier to work with the ***logarithm of the likelihood function*** (log-likelihood), denoted as $\ln L$.
- Taking the natural logarithm of L, we obtain the log-likelihood function:

$$\ln L(Y, \boldsymbol{\theta}, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y - X\boldsymbol{\theta})^T (Y - X\boldsymbol{\theta}).$$

Singular Case
Generalized Inverse
Estimability of LPF
MLE
Generalized LRM
GLS
Linear Constraints

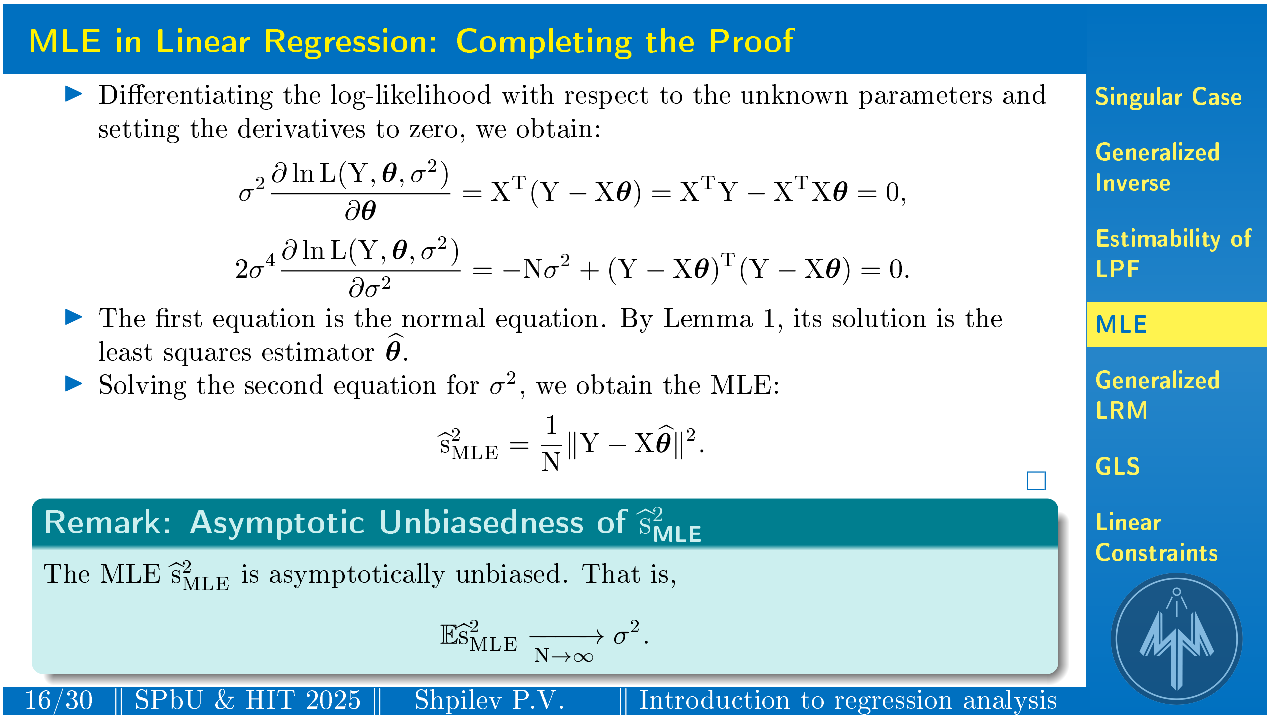
Comments

Let us consider the classical linear regression model under the standard assumptions on the error term: zero mean (unbiasedness), uncorrelated components, and homoskedasticity, meaning that all errors have the same variance. In addition to these standard conditions, we now impose a stronger assumption: we assume that the error vector follows a multivariate normal distribution.

Under these conditions, the observation vector Y has a normal distribution with mean $X\theta$ and covariance matrix $\sigma^2 I$. In this setting, the likelihood function can be written down explicitly, and we can formulate the following result: the maximum likelihood estimator for the parameter vector θ coincides with the usual least squares estimator. The estimator for the variance parameter σ^2 is given by $\frac{1}{N}$ times the squared norm of the residual vector $Y - X\widehat{\theta}$, evaluated at the estimated value of θ. This quantity is typically denoted by $\widehat{s}^2_{MLE}$.

To derive this result, we maximize the likelihood function with respect to the unknown parameters — namely, the vector θ and the scalar σ^2. Since the likelihood depends on $Y - X\theta$, the essential part of the optimization reduces to minimizing the corresponding quadratic form.

Rather than maximizing the likelihood directly, we take the natural logarithm. This simplifies the expression while preserving the maximizers. The resulting log-likelihood separates into two terms: the first involves the logarithm of σ^2, and the second involves the squared norm of $Y - X\theta$, scaled by σ^2. The structure of the log-likelihood makes clear the equivalence with least squares in θ and yields an explicit formula for the MLE of σ^2, written as $\frac{1}{N}$ times the squared norm of the residual vector.



MLE in Linear Regression: Completing the Proof

- Differentiating the log-likelihood with respect to the unknown parameters and setting the derivatives to zero, we obtain:

$$\sigma^2 \frac{\partial \ln \mathrm{L}(\mathrm{Y}, \boldsymbol{\theta}, \sigma^2)}{\partial \boldsymbol{\theta}} = \mathrm{X}^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\boldsymbol{\theta}) = \mathrm{X}^{\mathrm{T}}\mathrm{Y} - \mathrm{X}^{\mathrm{T}}\mathrm{X}\boldsymbol{\theta} = 0,$$

$$2\sigma^4 \frac{\partial \ln \mathrm{L}(\mathrm{Y}, \boldsymbol{\theta}, \sigma^2)}{\partial \sigma^2} = -\mathrm{N}\sigma^2 + (\mathrm{Y} - \mathrm{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\boldsymbol{\theta}) = 0.$$

- The first equation is the normal equation. By Lemma 1, its solution is the least squares estimator $\widehat{\boldsymbol{\theta}}$.
- Solving the second equation for σ^2, we obtain the MLE:

$$\widehat{\mathrm{s}}^2_{\mathrm{MLE}} = \frac{1}{\mathrm{N}}\|\mathrm{Y} - \mathrm{X}\widehat{\boldsymbol{\theta}}\|^2.$$

□

Remark: Asymptotic Unbiasedness of $\widehat{\mathrm{s}}^2_{\mathrm{MLE}}$

The MLE $\widehat{\mathrm{s}}^2_{\mathrm{MLE}}$ is asymptotically unbiased. That is,

$$\mathbb{E}\widehat{\mathrm{s}}^2_{\mathrm{MLE}} \xrightarrow[\mathrm{N}\to\infty]{} \sigma^2.$$

Singular Case

Generalized Inverse

Estimability of LPF

MLE

Generalized LRM

GLS

Linear Constraints

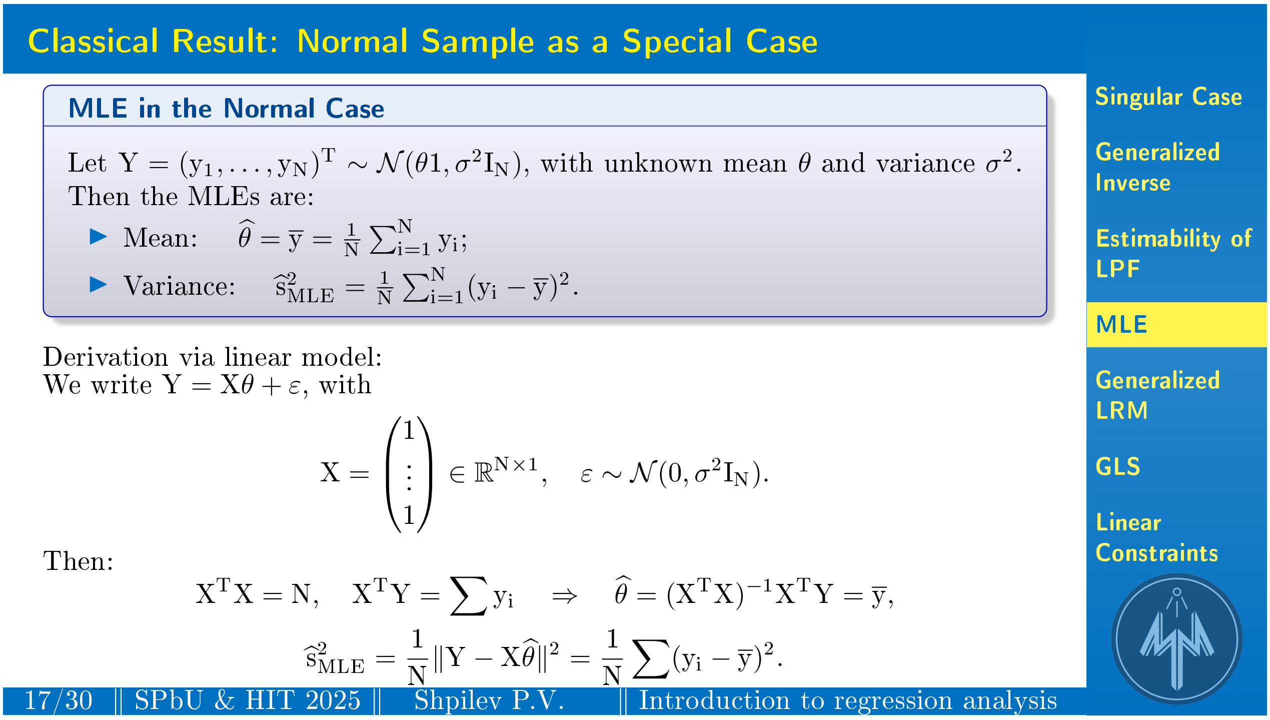
Comments

To complete the proof, we compute the partial derivatives of the log-likelihood with respect to the unknown parameters — the vector θ and the scalar σ^2 — and set them equal to zero. The derivative with respect to θ yields the matrix equation: σ^2 times the gradient equals the transpose of matrix X multiplied by vector $\mathrm{Y} - \mathrm{X}\theta$. Simplifying, we obtain the normal equations: $\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta = \mathrm{X}^{\mathrm{T}}\mathrm{Y}$. By Lemma 1, the unique solution of this system, provided X has full rank, is the ordinary least squares estimator of θ.

Next, we differentiate with respect to σ^2. After simplification and clearing denominators, we find that $2\sigma^4$ times the derivative equals $-\mathrm{N}\sigma^2$ plus the squared norm of the residual vector. Solving this equation for σ^2 yields the maximum likelihood estimator: $\widehat{\mathrm{s}}^2_{\mathrm{MLE}} = \frac{1}{\mathrm{N}}\|\mathrm{Y} - \mathrm{X}\widehat{\theta}\|^2$.

The theorem is thus fully proven.

Finally, we note that although this estimator for σ^2 is biased in finite samples, it becomes unbiased in the limit. That is, the expectation of $\widehat{\mathrm{s}}^2_{\mathrm{MLE}}$ converges to σ^2 as N tends to infinity. This property, known as asymptotic unbiasedness, justifies using this estimator in large-sample contexts.



Classical Result: Normal Sample as a Special Case

MLE in the Normal Case

Let $Y = (y_1, \dots, y_N)^T \sim \mathcal{N}(\theta 1, \sigma^2 I_N)$, with unknown mean θ and variance σ^2. Then the MLEs are:

- Mean: $\widehat{\theta} = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$;
- Variance: $\widehat{s}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$.

Derivation via linear model:
We write $Y = X\theta + \varepsilon$, with

$$X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{N\times 1}, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_N).$$

Then:

$$X^T X = N, \quad X^T Y = \sum y_i \quad \Rightarrow \quad \widehat{\theta} = (X^T X)^{-1} X^T Y = \bar{y},$$

$$\widehat{s}^2_{MLE} = \frac{1}{N}\|Y - X\widehat{\theta}\|^2 = \frac{1}{N}\sum (y_i - \bar{y})^2.$$

Comments

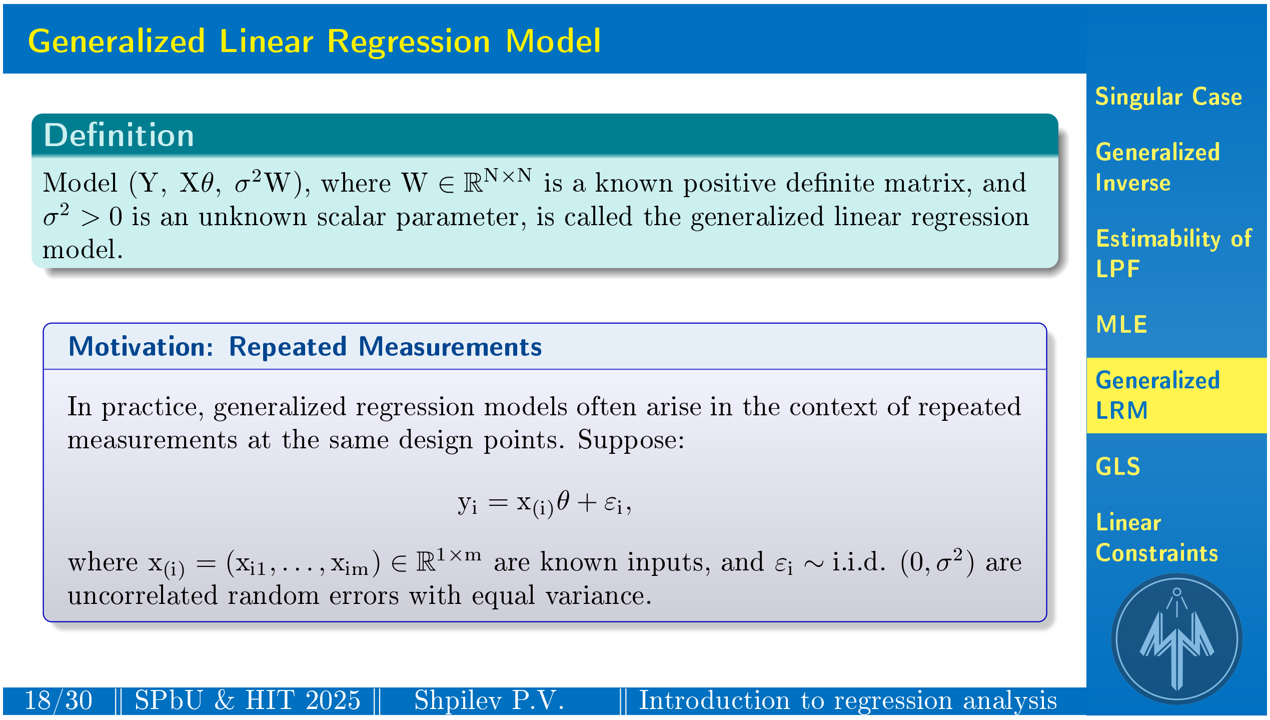
As an immediate corollary of Theorem Five, we obtain a classical result from mathematical statistics. Let us consider a one-dimensional repeated sample from a normal distribution with unknown mean θ and unknown variance σ^2. The maximum likelihood estimator of θ in this case is the sample mean, and the maximum likelihood estimator of σ^2 is the uncorrected sample variance, that is, the sum of squared deviations from the mean divided by N.

To formalize this, we write the sample as a linear regression model. The response vector Y consists of the values y_1 through y_N. The design matrix X consists of a single column of ones. Then the model $Y = X\theta + \epsilon$ corresponds to assuming all observations are identically distributed with common mean θ and independent, homoscedastic normal errors.

In this setting, the product of X^T and X is simply N, and the product of X^T and Y is the sum of all sample values. Therefore, the MLE for θ, which equals the inverse of $X^T X$ times $X^T Y$, reduces to the sample mean.

By applying the general formula for the MLE of σ^2, we substitute the estimated θ and obtain the uncorrected sample variance: $\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$.

This classical example illustrates how the general theory of linear models includes basic parametric estimation problems as special cases.



Generalized Linear Regression Model

Definition

Model $(Y,\ X\theta,\ \sigma^2 W)$, where $W \in \mathbb{R}^{N\times N}$ is a known positive definite matrix, and $\sigma^2 > 0$ is an unknown scalar parameter, is called the generalized linear regression model.

Motivation: Repeated Measurements

In practice, generalized regression models often arise in the context of repeated measurements at the same design points. Suppose:

$$y_i = x_{(i)}\theta + \varepsilon_i,$$

where $x_{(i)} = (x_{i1}, \dots, x_{im}) \in \mathbb{R}^{1\times m}$ are known inputs, and $\varepsilon_i \sim$ i.i.d. $(0, \sigma^2)$ are uncorrelated random errors with equal variance.

Comments

We now extend the classical linear regression model to a more general setting. Consider the model where the response vector Y has expectation equal to $X\theta$ and covariance matrix equal to $\sigma^2 W$, where W is a known symmetric positive definite matrix of size $N \times N$. This model is called the generalized linear regression model.

The key difference from the classical linear model is that the covariance matrix of the errors is no longer proportional to the identity. Instead, it is a general known matrix W, while σ^2 remains an unknown scalar. This structure allows modeling of heteroskedasticity and correlation among observations.

A common practical situation that leads to this model is when repeated measurements are taken at the same set of design points. Suppose that each observation y_i is generated according to the linear relation $y_i = x_i^T\theta + \epsilon_i$, where x_i is a known row vector and ϵ_i is a random error with variance σ^2. If we have multiple measurements at the same point, the errors across observations may be uncorrelated but not identically distributed, leading to a non-scalar covariance structure.

This motivates the use of a general weight matrix W in the model. In such cases, generalized least squares methods are appropriate for estimation.

Modeling Repeated Observations: The Averaged Approach

We have N total observations $x_{(i)}$, but only M are ***distinct*** points $t_{(1)}, \dots, t_{(M)}$. Let r_k be the ***number of measurements*** at each distinct point $t_{(k)}$ ($\sum_{k=1}^{M} r_k = N$).

- By ***averaging observations*** (y_j) at each $t_{(i)}$, we form a new model for $\widetilde{y}_i$:
$$\widetilde{y}_i = t_{(i)}\boldsymbol{\theta} + \widetilde{\epsilon}_i, \quad i = 1, \dots, M.$$
- Here, $\widetilde{y}_i$ are mean observations, $\widetilde{\epsilon}_i$ are mean errors.

Averaged Error Properties and Generalized Model Form

- **Error Properties:** $\mathbb{E}[\widetilde{\epsilon}_i] = 0$, $\mathbb{E}[\widetilde{\epsilon}_i\widetilde{\epsilon}_j] = 0$ for $i \neq j$.
- **Scaled Variance:** $\mathbb{E}[\widetilde{\epsilon}_i^2] = \frac{\sigma^2}{r_i}$.
- This leads to a ***Generalized Linear Regression Model***:
$$\left(\widetilde{Y}, \widetilde{X}\boldsymbol{\theta}, \sigma^2 W\right), \quad W = \operatorname{diag}\left(\frac{1}{r_1}, \dots, \frac{1}{r_M}\right).$$
- ***Key Implication:*** This model structure directly implies the need for ***Weighted Least Squares*** (WLS) for efficient estimation.

Comments

In many experimental settings, repeated measurements are collected at a limited number of design points. Suppose that among the N total measurements, only M design points are distinct. Denote these distinct points by t_1 through t_M. Let r_k denote the number of repeated observations taken at point t_k. Then the total number of measurements equals the sum over k from one to M of r_k, which is equal to N.

By averaging all measurements taken at each design point, we can reduce the original model to a new aggregated model with only M observations. This aggregated model has the form: $\widetilde{y}_i = t_i\theta + \widetilde{\epsilon}_i$, for i from one to M, where $\widetilde{y}_i$ is the average of all measurements at point t_i, and $\widetilde{\epsilon}_i$ is the corresponding average of the error terms.

Because the original errors were uncorrelated and had equal variance σ^2, the aggregated errors also remain uncorrelated, and their variances become σ^2/r_i. Hence, the new model has uncorrelated errors with unequal variances.

The result is a generalized linear regression model with response vector $\widetilde{Y}$, design matrix $\widetilde{X}$, and a diagonal covariance matrix W, where the i-th diagonal entry is $1/r_i$. The covariance structure of the noise becomes $\sigma^2 W$.

Remark on Model Equivalence: Generalized to Classical

Remark: Model Equivalence

A ***generalized linear regression model*** $(\mathrm{Y}, \mathrm{X}\theta, \sigma^2\mathrm{W})$ is equivalent to a ***classical linear regression model*** $(\widetilde{\mathrm{Y}}, \widetilde{\mathrm{X}}\theta, \sigma^2\mathrm{I}_\mathrm{N})$.

Justification for the Remark

- Start with generalized model $\big(\mathrm{Y}, \mathrm{X}\boldsymbol{\theta}, \sigma^2\mathrm{W}\big)$.
- Since W is ***positive definite***, $\exists$ non-singular A s.t. $\mathrm{W} = \mathrm{A}\mathrm{A}^\mathrm{T}$.
- ***Transform variables*** by A^{-1}: $\widetilde{\mathrm{Y}} = \mathrm{A}^{-1}\mathrm{Y}$, $\widetilde{\mathrm{X}} = \mathrm{A}^{-1}\mathrm{X}$, $\widetilde{\boldsymbol{\epsilon}} = \mathrm{A}^{-1}\boldsymbol{\epsilon}$.
- The ***covariance of transformed errors*** becomes:

$$\begin{aligned} \mathrm{D}(\widetilde{\boldsymbol{\epsilon}}) &= \mathbb{E}[\mathrm{A}^{-1}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\mathrm{T}(\mathrm{A}^{-1})^\mathrm{T}] \\ &= \mathrm{A}^{-1}(\sigma^2\mathrm{W})(\mathrm{A}^{-1})^\mathrm{T} \\ &= \sigma^2\mathrm{A}^{-1}(\mathrm{A}\mathrm{A}^\mathrm{T})(\mathrm{A}^{-1})^\mathrm{T} = \sigma^2\mathrm{I}. \end{aligned}$$

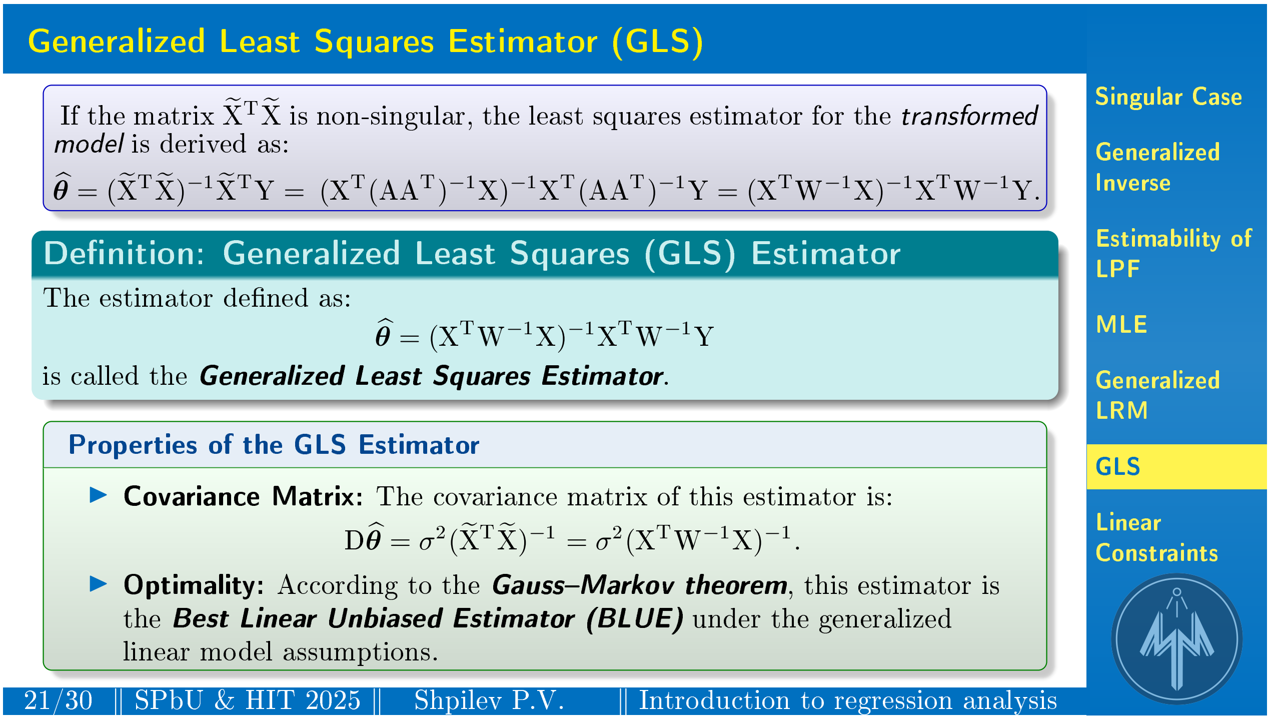
- This confirms the transformed model $\Big(\widetilde{\mathrm{Y}}, \widetilde{\mathrm{X}}\boldsymbol{\theta}, \sigma^2\mathrm{I}_\mathrm{N}\Big)$ has ***homoscedastic*** and ***uncorrelated errors***, acting as a classical regression model.

Comments

This slide establishes an equivalence between the generalized and classical linear regression models. We consider a generalized model where the error covariance matrix is $\sigma^2\mathrm{W}$, a known positive definite matrix. According to the spectral factorization theorem, such a matrix W can be written as the product $\mathrm{A}\mathrm{A}^\mathrm{T}$, where A is an invertible matrix.

We then transform the model by multiplying both sides by the inverse of A. This gives us a new response vector, design matrix, and error vector — denoted by $\widetilde{\mathrm{Y}}$, $\widetilde{\mathrm{X}}$, and $\widetilde{\epsilon}$. The transformed error term becomes $\widetilde{\epsilon} = \mathrm{A}^{-1}\epsilon$. Since ϵ had covariance $\sigma^2\mathrm{W}$, the new error has covariance $\sigma^2\mathrm{I}$. This transforms the model into the classical regression form, where the errors are independent and identically distributed with constant variance.

This result is highly useful in both theory and applications. It shows that every generalized linear model can be reduced to a classical one via an appropriate linear transformation, without changing the underlying structure of the model or the parameter of interest.



Generalized Least Squares Estimator (GLS)

If the matrix $\widetilde{X}^T\widetilde{X}$ is non-singular, the least squares estimator for the *transformed model* is derived as:

$$\widehat{\boldsymbol{\theta}} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T Y = (X^T(AA^T)^{-1}X)^{-1}X^T(AA^T)^{-1}Y = (X^TW^{-1}X)^{-1}X^TW^{-1}Y.$$

Definition: Generalized Least Squares (GLS) Estimator

The estimator defined as:

$$\widehat{\boldsymbol{\theta}} = (X^TW^{-1}X)^{-1}X^TW^{-1}Y$$

is called the ***Generalized Least Squares Estimator***.

Properties of the GLS Estimator

- **Covariance Matrix:** The covariance matrix of this estimator is:

$$D\widehat{\boldsymbol{\theta}} = \sigma^2(\widetilde{X}^T\widetilde{X})^{-1} = \sigma^2(X^TW^{-1}X)^{-1}.$$

- **Optimality:** According to the ***Gauss–Markov theorem***, this estimator is the ***Best Linear Unbiased Estimator (BLUE)*** under the generalized linear model assumptions.

Singular Case
Generalized Inverse
Estimability of LPF
MLE
Generalized LRM
GLS
Linear Constraints

Comments

In this slide, we formally define the generalized least squares estimator and establish its properties. Recall that after transforming the generalized regression model with covariance matrix $\sigma^2 W$ into the classical form using a spectral factorization, we obtained a model with independent homoscedastic errors. In that classical model, the ordinary least squares estimator is given by the inverse of $\widetilde{X}^T\widetilde{X}$, multiplied by $\widetilde{X}^T\widetilde{Y}$.

Substituting $\widetilde{X}$ as $A^{-1}X$, and $\widetilde{Y}$ as $A^{-1}Y$, we express the estimator entirely in terms of the original variables X, Y, and the matrix W, which equals AA^T. After simplification, we arrive at the expression: $\widehat{\theta} = (X^TW^{-1}X)^{-1}X^TW^{-1}Y$. This is called the generalized least squares estimator.

It is important to note that this estimator explicitly depends on the covariance structure of the errors via the matrix W. Its variance-covariance matrix, as given on the slide, is $\sigma^2(X^TW^{-1}X)^{-1}$.

Finally, the Gauss–Markov theorem ensures that among all linear unbiased estimators, the generalized least squares estimator has the minimal variance. That is, it is the best linear unbiased estimator, or BLUE, for the parameter vector θ.

Regression with Prior Information on Parameters (Cases 1–2)

Motivation

In many practical problems, the researcher has prior knowledge about the model parameters. This can significantly improve estimation accuracy if appropriately incorporated.

Four Typical Types of Prior Information

We will consider four common situations where prior information is available:

1) **Exact Linear Constraints:**

   The parameters satisfy known linear equalities, such as $\mathrm{R}\boldsymbol{\theta} = \mathrm{u}$.

   This reduces to the classical constrained least squares problem.

2) **Noisy Prior Information:**

   The prior knowledge comes from previous measurements or a related model and takes the form:

$$\mathrm{u} = \mathrm{R}\theta + \zeta, \quad \zeta \sim (0, \mathrm{D}), \quad \mathrm{D} > 0, \quad \mathbb{E}[\zeta\epsilon^{\mathrm{T}}] = 0$$

   Here, $\boldsymbol{\zeta}$ is a random error vector independent of the model error $\boldsymbol{\varepsilon}$.

Comments

In many real-world applications, researchers are not working in complete uncertainty — they often possess some prior knowledge about the parameters of the regression model. Incorporating such prior information can improve estimation accuracy and model interpretability.

We will examine four common types of prior information. The first two are introduced here; the remaining two will be considered next.

Case 1 involves exact linear restrictions on the parameter vector — for example, you might know that certain linear combinations of parameters must equal fixed values. In this case, estimation reduces to the familiar problem of least squares with linear equality constraints, which we've already studied.

Case 2 is more nuanced: the researcher has prior data from similar experiments or past studies. This information may not be exact but is still informative. In this setup, the prior information is expressed as a random vector equation: "$\mathrm{u} = \mathrm{R}\theta +$noise". The noise term, denoted by ζ, is assumed to have some distribution with zero mean and a known positive definite covariance matrix D. It's also uncorrelated with the main model's random error.

This framework allows us to formally include uncertain, yet structured, information in our estimation process.

Regression with Prior Information on Parameters (Cases 3–4)

Singular Case
Generalized Inverse
Estimability of LPF
MLE
Generalized LRM
GLS
Linear Constraints

Four Typical Types of Prior Information (continued)

We now describe the remaining two types of prior information:

3) **Inequality Constraints:**
   The parameter vector lies within a known ellipsoid:
   $$\Omega = \left\{\boldsymbol{\theta} \in \mathbb{R}^{\mathrm{m}} \mid (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathrm{A} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \leq \mathrm{k}\right\}$$
   Here, $\boldsymbol{\theta}_0$ is the center, A is a positive definite matrix, and $\mathrm{k} > 0$ is the size parameter. This leads to a minimax estimation approach.
4) **Probabilistic Prior:**
   The parameter vector is assumed to have a known distribution within a region such as Ω. In this case, the problem reduces to classical generalized least squares.

Note. Each type of prior information leads to a different estimation principle: exact constraints yield constrained least squares; linear noisy priors lead to generalized least squares; ellipsoidal constraints require minimax estimators; and probabilistic priors often motivate Bayesian or regularized methods.

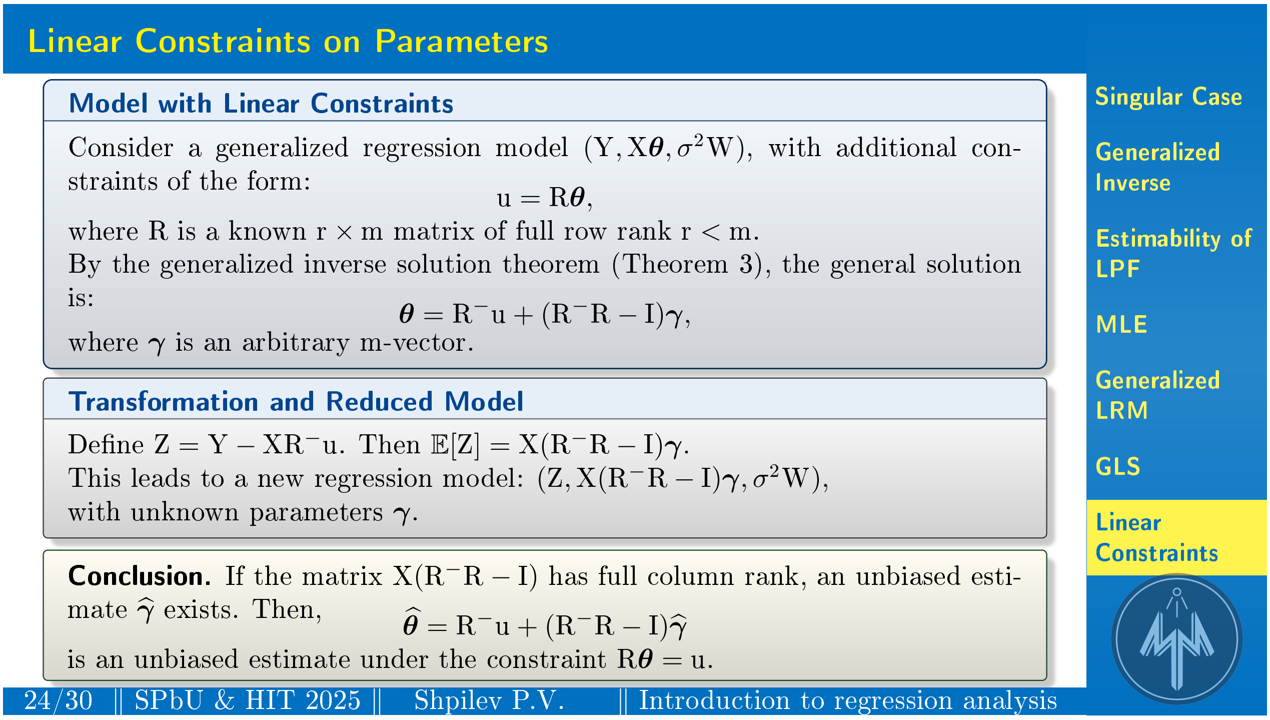
Comments

Case 3 involves inequality constraints on the parameter vector. A common example is when it is known that the parameters lie within a certain ellipsoid. This ellipsoid is defined as the set of all parameter vectors such that the quadratic form — $(\theta - \theta_0)^{\mathrm{T}} \mathrm{A}(\theta - \theta_0)$ — is less than or equal to a positive constant k. Here, θ_0 is the center of the ellipsoid, A is a positive definite matrix, and k determines the size. Under such conditions, standard estimation techniques are replaced by minimax procedures, which aim to guard against the worst-case scenario within the given region.

Case 4 reflects a probabilistic form of prior knowledge. In this case, the researcher knows not only that the parameter lies in a certain region, such as the ellipsoid Ω, but also assumes a full probability distribution for the parameter vector within that region. For example, the vector θ may follow a normal distribution centered at θ_0 with covariance matrix proportional to A^{-1}. In such cases, the estimation problem is equivalent to classical procedures such as the generalized least squares or Bayesian estimators, depending on the exact specification.

Let us emphasize: each of the four types of prior information leads to a distinct estimation approach. When constraints are exact equalities, we use constrained least squares. When the prior is linear with noise, generalized least squares applies. Ellipsoidal constraints require minimax estimators, while probabilistic priors motivate Bayesian or regularized methods. These distinctions help select appropriate tools for incorporating prior knowledge in regression analysis.

Let us now consider each of these cases in more detail.



Linear Constraints on Parameters

Model with Linear Constraints

Consider a generalized regression model $(\mathrm{Y}, \mathrm{X}\boldsymbol{\theta}, \sigma^2\mathrm{W})$, with additional constraints of the form:

$$\mathrm{u} = \mathrm{R}\boldsymbol{\theta},$$

where R is a known $\mathrm{r} \times \mathrm{m}$ matrix of full row rank $\mathrm{r} < \mathrm{m}$.
By the generalized inverse solution theorem (Theorem 3), the general solution is:

$$\boldsymbol{\theta} = \mathrm{R}^{-}\mathrm{u} + (\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\boldsymbol{\gamma},$$

where $\boldsymbol{\gamma}$ is an arbitrary m-vector.

Transformation and Reduced Model

Define $\mathrm{Z} = \mathrm{Y} - \mathrm{X}\mathrm{R}^{-}\mathrm{u}$. Then $\mathbb{E}[\mathrm{Z}] = \mathrm{X}(\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\boldsymbol{\gamma}$.
This leads to a new regression model: $(\mathrm{Z}, \mathrm{X}(\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\boldsymbol{\gamma}, \sigma^2\mathrm{W})$,
with unknown parameters $\boldsymbol{\gamma}$.

Conclusion. If the matrix $\mathrm{X}(\mathrm{R}^{-}\mathrm{R} - \mathrm{I})$ has full column rank, an unbiased estimate $\widehat{\boldsymbol{\gamma}}$ exists. Then,

$$\widehat{\boldsymbol{\theta}} = \mathrm{R}^{-}\mathrm{u} + (\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\widehat{\boldsymbol{\gamma}}$$

is an unbiased estimate under the constraint $\mathrm{R}\boldsymbol{\theta} = \mathrm{u}$.

Singular Case
Generalized Inverse
Estimability of LPF
MLE
Generalized LRM
GLS
Linear Constraints

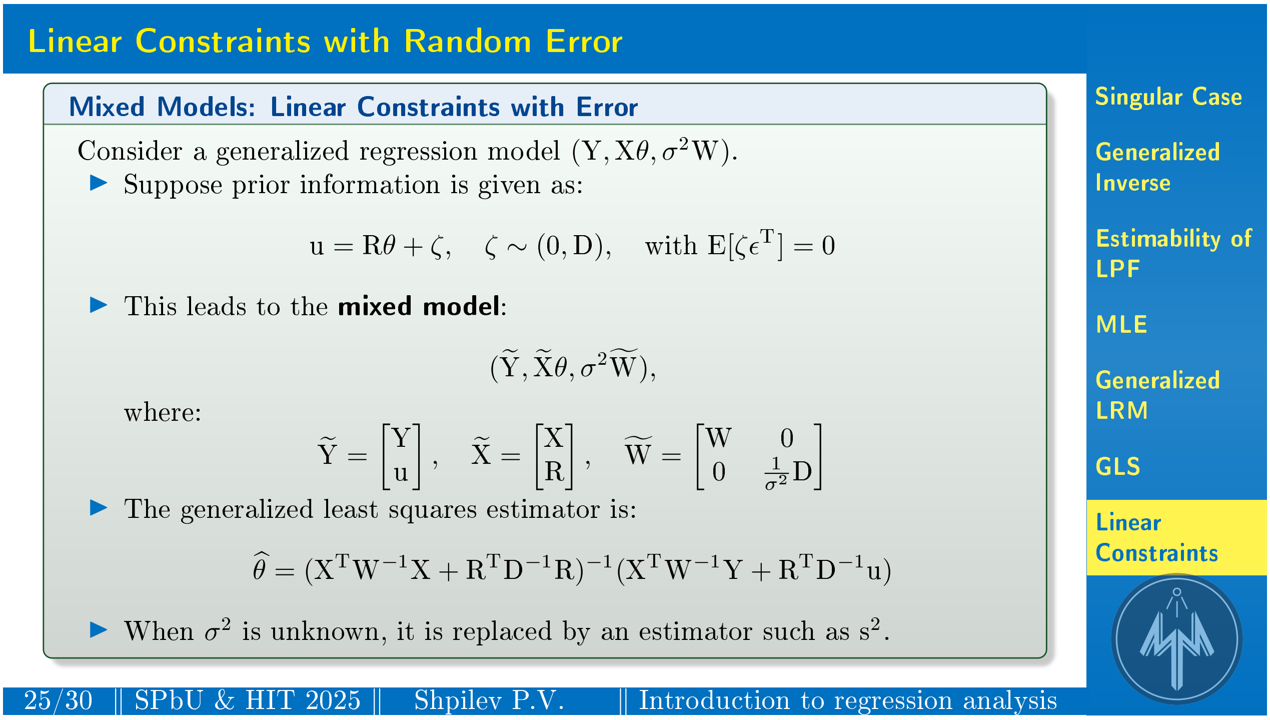
Comments

We now consider the case when the parameter vector in a generalized regression model is subject to linear constraints. Specifically, suppose we have a model with observations Y, design matrix X, parameter vector θ, and covariance matrix $\sigma^2\mathrm{W}$. Additionally, suppose that the parameters satisfy a known linear constraint of the form $\mathrm{u} = \mathrm{R}\theta$, where the matrix R has dimensions $\mathrm{r} \times \mathrm{m}$ and full rank r, with r strictly less than m.

According to the theorem on generalized inverse solutions, the general solution of the constraint equation is given by $\theta = \mathrm{R}^{-}\mathrm{u} + (\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\gamma$. Here, γ is an arbitrary m-dimensional vector.

We now substitute this expression into the original model. Let us define a new variable Z, equal to $\mathrm{Y} - \mathrm{X}\mathrm{R}^{-}\mathrm{u}$. Its expectation equals $\mathrm{X}(\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\gamma$. This allows us to consider a new regression model where Z is the new vector of observations, and the parameter to be estimated is γ.

If the new design matrix $\mathrm{X}(\mathrm{R}^{-}\mathrm{R} - \mathrm{I})$ has full column rank, then γ can be estimated without bias. Finally, we reconstruct an estimate for θ using the formula: $\widehat{\theta} = \mathrm{R}^{-}\mathrm{u} + (\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\widehat{\gamma}$. This estimator satisfies the constraint and is unbiased.



Linear Constraints with Random Error

Mixed Models: Linear Constraints with Error

Consider a generalized regression model $(\mathrm{Y}, \mathrm{X}\theta, \sigma^2\mathrm{W})$.

- Suppose prior information is given as:

$$\mathrm{u} = \mathrm{R}\theta + \zeta, \quad \zeta \sim (0, \mathrm{D}), \quad \text{with } \mathrm{E}[\zeta\epsilon^{\mathrm{T}}] = 0$$

- This leads to the **mixed model**:

$$(\widetilde{\mathrm{Y}}, \widetilde{\mathrm{X}}\theta, \sigma^2\widetilde{\mathrm{W}}),$$

where:

$$\widetilde{\mathrm{Y}} = \begin{bmatrix} \mathrm{Y} \\ \mathrm{u} \end{bmatrix}, \quad \widetilde{\mathrm{X}} = \begin{bmatrix} \mathrm{X} \\ \mathrm{R} \end{bmatrix}, \quad \widetilde{\mathrm{W}} = \begin{bmatrix} \mathrm{W} & 0 \\ 0 & \frac{1}{\sigma^2}\mathrm{D} \end{bmatrix}$$

- The generalized least squares estimator is:

$$\widehat{\theta} = (\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X} + \mathrm{R}^{\mathrm{T}}\mathrm{D}^{-1}\mathrm{R})^{-1}(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{Y} + \mathrm{R}^{\mathrm{T}}\mathrm{D}^{-1}\mathrm{u})$$

- When σ^2 is unknown, it is replaced by an estimator such as s^2.

Comments

We now consider the case when the prior information about parameters is inexact, expressed with random error. Suppose we have a generalized regression model where the response vector is Y, the design matrix is X, the parameter vector is θ, and the covariance matrix of errors is σ^2W. In addition, assume that some auxiliary measurement provides a vector u, related to θ via the equation $\mathrm{u} = \mathrm{R}\theta + \zeta$, where ζ is a random error vector with mean zero and covariance matrix D. This leads to what is called a mixed model, since it combines observation equations with noisy prior constraints.

This mixed model can be rewritten in an extended form with augmented data and design matrices. The extended response vector is the vertical concatenation of Y and u. The extended design matrix stacks X over R. The extended covariance matrix is block-diagonal, with W in the upper-left block and $\frac{1}{\sigma^2}\mathrm{D}$ in the lower-right block.

Using generalized least squares, we obtain the estimator of θ as the inverse of the sum of $\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}$ and $\mathrm{R}^{\mathrm{T}}\mathrm{D}^{-1}\mathrm{R}$, multiplied by the sum of $\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{Y}$ and $\mathrm{R}^{\mathrm{T}}\mathrm{D}^{-1}\mathrm{u}$. This estimator incorporates both the original observations and the prior information with uncertainty. In practice, the value of σ^2 is often unknown and must be replaced by an estimate, such as s^2 from Theorem 5.

Parameter Set Constraints: Minimax Estimation

Constraints by Region (Set-Based Information)

Let the generalized regression model be given as $(\mathrm{Y}, \mathrm{X}\theta, \sigma^2\mathrm{W})$, and assume that the parameter vector θ is known to lie in a specified region $\Omega \subset \mathbb{R}^{\mathrm{m}}$.

Definition: Minimax Estimator

The **minimax estimator** $\widehat{\theta}$ is defined by

$$\widehat{\theta} = \arg \min_{\widetilde{\theta}\in\Omega} \max_{\theta\in\Omega} \mathrm{g}(\theta, \widetilde{\theta}),$$

where

$$\mathrm{g}(\theta, \widetilde{\theta}) = \mathrm{a}^{\mathrm{T}}\,\mathrm{E}\left[(\widetilde{\theta} - \theta)(\widetilde{\theta} - \theta)^{\mathrm{T}}\right]\mathrm{a},$$

with $\widetilde{\theta}$ ranging over linear estimators of θ, and $\mathrm{a} \in \mathbb{R}^{\mathrm{n}}$ is any fixed nonzero vector.

When $\mathrm{a} = \mathrm{e}_{\mathrm{k}}$, the k-th standard basis vector, the corresponding component $\widehat{\theta}_{\mathrm{k}}$ is the best minimax linear estimator of the coordinate θ_{k}.

Comments

We now consider a generalized regression model where the vector of observations is Y, the design matrix is X, the parameter vector is θ, and the covariance matrix of the errors is $\sigma^2\mathrm{W}$. Suppose that, in addition to this model, prior information is available in the form of a constraint: the true parameter vector θ is known to belong to some fixed subset Ω of the m-dimensional space.

In such cases, a natural approach is to seek an estimator that is robust under worst-case conditions within Ω. This leads to the minimax principle. A minimax estimator of θ is defined as the linear estimator that minimizes the maximum expected loss over all possible true values of θ in Ω. Specifically, we define the estimator as the argument minimum over all $\widetilde{\theta}$ in Ω of the maximum, over θ in Ω, of the quantity $\mathrm{g}(\theta, \widetilde{\theta})$, where this quantity equals $\mathrm{a}^{\mathrm{T}}\mathrm{E}[(\widetilde{\theta} - \theta)(\widetilde{\theta} - \theta)^{\mathrm{T}}]\mathrm{a}$.

Here, the vector a is any fixed nonzero vector in n-dimensional space. In practice, one is often interested in estimating individual components of θ. If we take a to be the standard basis vector e_{k}, which has one in the k-th coordinate and zeros elsewhere, then the k-th component of the minimax estimator is the best linear estimator for the k-th component of θ in the minimax sense.

Minimax Estimation under Quadratic Region Constraints

Explicit Minimax Estimator under Set Constraints

Let the generalized regression model be given by $(Y, X\theta, \sigma^2 W)$, and suppose that the parameter vector θ is constrained to a set of the form

$$\Omega = \{\theta \in \mathbb{R}^m : (\theta - \theta_0)^T A(\theta - \theta_0) \le k\},$$

where A is a positive definite matrix and $k > 0$.

Theorem 6

Under the above model and constraint, the minimax estimator has the explicit form

$$\widehat{\theta}_{mM} = \left(\frac{\sigma^2}{k}A + X^T W^{-1} X\right)^{-1} X^T W^{-1}(Y - X\theta_0) + \theta_0.$$

A detailed proof of this result can be found in **Ermakov, S. M., Zhiglyavsky, A. A. (1987). Mathematical Theory of Optimal Experiment. Moscow: Nauka** (Theorem 2.1, page 42). The formula also shows that, for sets Ω of the given quadratic form, the minimax estimator $\widehat{\theta}_{mM}$ does not depend on the choice of vector **a** from the minimax definition.

Comments

We now present an explicit result describing the minimax estimator in the case where the parameter vector θ is known to lie within a closed ellipsoidal region. This region is defined as the set of all θ such that the quadratic form $(\theta - \theta_0)^T A(\theta - \theta_0)$ is less than or equal to k. The matrix A is symmetric and positive definite, and the constant k is strictly positive.

Under these assumptions, the minimax estimator, which minimizes the worst-case quadratic risk within this region, has a closed-form expression. Specifically, the minimax estimator $\widehat{\theta}_{mM}$ is equal to the inverse of the matrix $\frac{\sigma^2}{k}A + X^T W^{-1} X$, multiplied by the vector $X^T W^{-1}(Y - X\theta_0)$, and then the result is shifted by θ_0.

This result is formally stated in the theorem and can be found in the book by Ermakov - Zhiglyavsky, page forty-two. Importantly, this formula reveals that, for regions of the given quadratic type, the minimax estimator no longer depends on the choice of the vector a appearing in the general definition of minimaxity. That is, the entire estimator becomes intrinsic to the constraint region, independent of which linear functional of θ we are focusing on.

Special Cases of the Minimax Estimator

Canonical Ellipsoid, Classical Model:
If the ellipsoid is centered at the origin and the model is classical, that is, $\theta_0 = 0$ and $\mathrm{W} = \mathrm{I_N}$, then the minimax estimator takes the form

$$\widehat{\theta}_{\mathrm{mM}} = \left(\frac{\sigma^2}{\mathrm{k}}\mathrm{A} + \mathrm{X^T X}\right)^{-1} \mathrm{X^T Y}.$$

Ridge Estimation Case:
If $\Omega = \{\theta : \theta^{\mathrm{T}}\theta \leq \sigma^2/\mathrm{k}\}$, then the minimax estimator becomes

$$\widehat{\theta}_{\mathrm{mM}} = \left(\mathrm{kI_m} + \mathrm{X^T X}\right)^{-1} \mathrm{X^T Y},$$

which is known as the ridge estimator.

Limiting Behavior: As $\mathrm{k} \to \infty$, the minimax estimator converges to the generalized least squares estimator. In other words, the GLS estimator is minimax in the absence of prior information.
Furthermore, for any finite k, the minimax estimator is always uniquely defined, even when the matrix $\mathrm{X^T X}$ is singular. This contrasts with the ordinary least squares estimator, which may not be uniquely defined in degenerate cases.

Singular Case
Generalized Inverse
Estimability of LPF
MLE
Generalized LRM
GLS
Linear Constraints

Comments

Let us consider three notable special cases of the minimax estimator. First, suppose that the center of the ellipsoid is at the origin and the regression model is classical, meaning that $\theta_0 = 0$ and the matrix W is the identity matrix of size N. Then the general minimax formula simplifies to the matrix inverse of $\frac{\sigma^2}{\mathrm{k}}\mathrm{A} + \mathrm{X^T X}$, multiplied by $\mathrm{X^T Y}$. This is a natural shrinkage form centered at zero.

Second, in the specific case where the constraint set Ω is the ball defined by all θ such that the scalar product $\theta^{\mathrm{T}}\theta$ is less than or equal to σ^2/k, the minimax estimator becomes the inverse of $\mathrm{kI_m} + \mathrm{X^T X}$, multiplied by $\mathrm{X^T Y}$. This is precisely the ridge estimator — a well-known regularized version of the least squares estimator.

Third, let us examine the limiting behavior. As the value of k tends to infinity, the minimax estimator tends to the generalized least squares estimator. This reflects the fact that when there is no prior information, the GLS estimator is minimax. Moreover, for any finite k, the minimax estimator is always well-defined, regardless of whether the matrix $\mathrm{X^T X}$ is invertible. This is a significant advantage over the classical least squares estimator, which requires resorting to generalized inverse matrices when the matrix $\mathrm{X^T X}$ is singular.

Minimax Property in the Class of Unbiased Estimators

Remark

The minimax estimator is generally biased. In the class of unbiased estimators, the generalized least squares (GLS) estimator is minimax.

Key idea: If $\widetilde{\theta} = \mathrm{C}\mathrm{Y}$ is unbiased, then $\mathrm{C}\mathrm{X} = \mathrm{I}_\mathrm{m}$. Therefore:

$$\widetilde{\theta} - \theta = \mathrm{C}\mathrm{Y} - \theta = \mathrm{C}(\mathrm{X}\theta + \varepsilon) - \theta = (\mathrm{C}\mathrm{X} - \mathrm{I}_\mathrm{m})\theta + \mathrm{C}\varepsilon = \mathrm{C}\varepsilon$$

- The estimation error $\mathrm{C}\varepsilon$ does not depend on θ.
- The risk function becomes:

$$\mathrm{g}(\theta, \widetilde{\theta}) = \sigma^2 \mathrm{a}^\mathrm{T}\mathrm{C}\mathrm{W}\mathrm{C}^\mathrm{T}\mathrm{a} = \mathrm{a}^\mathrm{T}\mathrm{D}_{\widetilde{\theta}}\mathrm{a}$$

- Thus, g does not depend on θ and is minimized over C.

Conclusion

By the Gauss–Markov theorem, the minimum is attained at:

$$\mathrm{C} = (\mathrm{X}^\mathrm{T}\mathrm{W}^{-1}\mathrm{X})^{-1}\mathrm{X}^\mathrm{T}\mathrm{W}^{-1}$$

Therefore, the GLS estimator is minimax in the class of unbiased estimators.

Comments

Let us emphasize that the general minimax estimator is, in general, a biased estimator. However, if we restrict attention to the class of unbiased estimators, then the generalized least squares estimator is minimax within this class. This conclusion follows directly from the requirement that any unbiased linear estimator of the form CY must satisfy the constraint that matrix C times matrix X equals the identity matrix of size m.

Under this constraint, the estimation error, which is the vector $\mathrm{C}\epsilon$, no longer depends on the true value of θ. Consequently, the mean squared error expression, which equals $\sigma^2\mathrm{a}^\mathrm{T}\mathrm{C}\mathrm{W}\mathrm{C}^\mathrm{T}\mathrm{a}$, is constant with respect to θ. Then, by the Gauss–Markov theorem for the generalized linear model, the minimum of this error expression is achieved when $\mathrm{C} = (\mathrm{X}^\mathrm{T}\mathrm{W}^{-1}\mathrm{X})^{-1}\mathrm{X}^\mathrm{T}\mathrm{W}^{-1}$.

That is, when the estimator is the generalized least squares estimator. Therefore, the GLS estimator is minimax among all unbiased linear estimators.

Bayesian Estimator with Known Prior Distribution

Prior Information

Assume a prior distribution $\mathrm{P}(d\theta)$ is given on a measurable space $(\Omega, \mathcal{F})$, independent of the noise vector ε, with:

- Mean vector: u
- Covariance matrix: D

This is equivalent to the prior constraint

$$\theta = \mathrm{u} + \zeta, \quad \zeta \sim (0, \mathrm{D}), \quad \mathrm{D} > 0, \quad \mathbb{E}[\zeta \epsilon^{\mathrm{T}}] = 0$$

Definition (Bayesian Estimator)

The posterior mean of θ under the prior $\mathrm{P}(d\theta)$ is given by:

$$\widehat{\theta} = (\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X} + \sigma^2\mathrm{D}^{-1})^{-1}(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{Y} + \sigma^2\mathrm{D}^{-1}\mathrm{u})$$

This is called the Bayesian estimator.

Comments

Let us now consider the case where prior information about the parameter vector θ is specified in the form of a known probability distribution. Concretely, we assume that there exists a prior distribution $\mathrm{P}(d\theta)$, defined on a sigma-algebra of subsets of a sample space Ω. This prior is independent of the distribution of the error vector ϵ.

The prior distribution is assumed to have mean vector u and covariance matrix D. This setup corresponds to a situation where the parameter vector θ satisfies a stochastic constraint of the form: $\theta = \mathrm{I}\theta + \zeta$, where ζ follows a normal distribution with mean zero and covariance matrix D.

Under these assumptions, we construct an estimator that minimizes the average prediction error, taking into account both the randomness in the data and the uncertainty in the parameter values. This leads us to the so-called Bayes estimator. The criterion being minimized is the expected squared distance between the estimator and the true parameter vector, where the expectation is taken over both the data and the prior distribution of θ.

The resulting estimator is called the posterior mean. It is computed as follows: the inverse of the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X} + \sigma^2\mathrm{D}^{-1}$, multiplied by the vector $\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{Y} + \sigma^2\mathrm{D}^{-1}\mathrm{u}$.

This estimator is sometimes referred to as a shrinkage estimator, because it balances between the data-driven estimate and the prior mean u.