PART III. Optimal design theory
(LECTURE 1)

Shpilev Petr Valerievich
Faculty of Mathematics and Mechanics, SPbU

September, 2025

Санкт-Петербургский государственный университет

Comments

The third part of our course introduces the principles of regression modeling and optimal experimental design. Many optimization problems rely on expensive function evaluations. To address this, we use surrogate models: smooth, low-cost approximations of the true objective function. These models are fast to evaluate and optimize, guiding us toward the true optimum efficiently. By some sampling of the real function, we refine the surrogate, improving its accuracy over time. Beyond optimization, this approach is powerful for descriptive regression, where we study how variables interact. It helps uncover relationships between inputs (e.g., material properties or hyperparameters) and outputs (e.g., performance or efficiency)—without strict assumptions about data distributions. So the third part of our course dives into regression analysis basics and optimal experiment design – core techniques for building and refining models.

In today's lecture, we introduce the foundational principles of descriptive regression and its application in optimal design theory. We begin by exploring the core idea of descriptive regression, with a focus on empirical data and parametric regression models. A key example is the modeling of height as a function of weight, including a visual representation and the process of estimating the model parameters. We examine the criteria for optimal parameter estimation and the importance of the least squares estimator (LSE), which leads us into the discussion of normal equations and classical linear regression models.

The lecture progresses with a detailed explanation of the Gauss-Markov theorem and the conditions for the Best Linear Unbiased Estimator (BLUE), illustrating the optimality of the LSE in terms of variance minimization. The proof of important lemmas, such as those related to the properties of the OLS estimator and variance of linear unbiased estimators, is presented in-depth. Through specific examples, we further explore the practical application of least squares estimation, both with exact data and in the context of quadratic approximation.

We conclude the lecture by introducing the concept of estimability of linear parametric functions, explaining the conditions under which these estimators are feasible and optimal. The lecture sets the stage for the more advanced topics in optimal design theory that follow.

Descriptive Regression: The Core Idea

Regression analysis is a statistical method for estimating the relationships among variables. It describes how a dependent variable changes as one or more independent variables change.

- **Purpose:** To represent and understand the relationship between a response (output) and several influencing factors (inputs).
- **Focus of Descriptive Regression:** To build a model based on observed data without making strong assumptions about the underlying statistical distribution.

Key Concepts

- **Dependent Variable (y):** The outcome or response variable we are trying to predict or explain.
- **Independent Variables ($x_1, \ldots, x_m$):** The predictor variables or factors that influence the dependent variable. Also called regressors.
- **Model Function ($y = \eta(x_1, \ldots, x_m)$):** A mathematical representation of the relationship.

Comments

Many optimization problems involve expensive function evaluations. For instance, testing a hardware design might require hours of fabrication, an aircraft design needs costly wind tunnel tests, and tuning deep learning hyperparameters could take a week of GPU training. To tackle this, we use surrogate models—smooth, inexpensive approximations of the true objective function. These models are quick to evaluate and optimize, guiding us toward the true optimum without excessive cost. By occasionally evaluating the real function, we can refine the surrogate model, making it more accurate over time. Beyond optimization, this approach is powerful for descriptive regression, which studies how variables interact. It helps us understand relationships between inputs, like material properties or hyperparameters, and outputs, like performance or efficiency, without assuming specific data distributions.

Let's give a formal description of this approach. Imagine we have a system that works like a black box. We input a set of control signals, say x_1 through x_m, which in practice are typically numerical quantities. The output is a single scalar value, y, which depends on x_1 through x_m. Our goal is to determine this relationship, that is, to find a function where y equals η of x_1 through x_m. To solve this task, the first step is to collect experimental data, meaning the results of simultaneous measurements of y and x_1 through x_m.

Empirical Data in Descriptive Regression

Setup: After conducting N experiments, we collect data to model the relationship between inputs and outputs.

- **Empirical Data Matrix:**

$$(Y, X) = \begin{bmatrix} y_1 & x_{11} & \cdots & x_{1m} \\ y_2 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_N & x_{N1} & \cdots & x_{Nm} \end{bmatrix}$$

Output Vector (Y): $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$ Design Matrix (X): $\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nm} \end{bmatrix}$

Key Terms:

Vector of Results (Y): Column of observed output values y_j for $j = 1, \ldots, N$.
Design Matrix (X): Also called plan matrix, stores inputs x_{ij}, often predefined.
Plan (design): Inputs x_{ij} can be set or measured before experiments.

Comments

This slide formalizes how we organize experimental data for regression analysis. After conducting N experiments, we structure our observations into two key components: the output vector Y and design matrix X. The output vector contains all observed values of our dependent variable y, while the design matrix systematically records all corresponding input variable values.

The design matrix plays a crucial role - its structure determines what relationships we can detect between inputs and outputs. Each row represents a complete experimental observation, while columns correspond to different input variables. This arrangement allows us to efficiently analyze multivariate relationships.

The terminology is important here: we distinguish between the "vector of results" (our measured outcomes) and the "design matrix" (our controlled or observed inputs). In many experimental setups, the x-values in the design matrix are carefully chosen in advance (the experimental "design"), while the y-values are measured outcomes that may contain noise. This structured approach enables us to apply mathematical and statistical tools to uncover the underlying relationship η between inputs and outputs.

Parametric Regression Models

Core Challenge: Exact functional relationships are often too complex to determine, so we use simplified parametric models that approximate the statistical dependence with satisfactory accuracy.

Parametric Regression Approach

- Choose a model function $\eta(\mathrm{x}, \theta)$ with parameters θ
- Find θ that best fits the observed data
- Linear vs nonlinear in parameters
- Balance between simplicity and accuracy

Model Types

Linear Models:

$$\eta(\mathrm{x}, \theta) = \theta_1 \mathrm{x}_1 + \cdots + \theta_\mathrm{m} \mathrm{x}_\mathrm{m}$$

With Intercept:

$$\eta(\mathrm{x}, \theta) = \theta_0 + \theta_1 \mathrm{x}_1 + \cdots + \theta_{\mathrm{m}-1} \mathrm{x}_{\mathrm{m}-1}$$

Nonlinear Models:

$\eta(\mathrm{x}, \theta) =$ Any non-linear function

Why Approximate?

- Real-world relationships are complex
- Measurement noise and variability
- Computational tractability
- Interpretability trade-off

Comments

Parametric regression provides a practical framework for modeling complex relationships when exact functional forms are unknown. The key idea is to select a family of functions defined by parameters θ that can approximate the true relationship with sufficient accuracy. Linear models are particularly important due to their simplicity and interpretability - they assume the output is a weighted sum of inputs. The inclusion of an intercept term (θ_0) accounts for baseline effects. More complex nonlinear models can capture intricate patterns but require careful handling to avoid overfitting. The choice between model complexity and simplicity involves trade-offs: simpler models are more robust and interpretable but may miss important relationships, while complex models can fit data better but may capture noise rather than signal. Ultimately, the goal is to find the sweet spot where the model is complex enough to be useful but simple enough to be reliable.

Modeling Height as a Function of Weight

Objective: Investigate the relationship between height (y) and weight (x_1) in a sample of adult men.

- The dataset consists of 35 points (x_{1j}, y_j), where x_1 is weight (kg) and y is height (cm).
- There is no strict functional dependence between x_1 and y: multiple weights can correspond to the same height and vice versa.
- However, the average trend of height as a function of weight can be modeled approximately.

Linear Approximation Model

$$y \approx \theta_0 + \theta_1 x_1$$

This simple linear form is widely used in practice to estimate statistical relationships.

Visualization: Bivariate plots (like the one shown next) help identify trends and statistical dependencies between variables.

Comments

Let's consider a simple but revealing example — modeling height as a function of weight.

Now, intuitively, we expect that taller people tend to weigh more. But does that mean there's a strict mathematical formula linking the two? Not quite. If you've ever compared athletes of the same weight — say, a basketball player and a swimmer — you'll know that height can vary a lot. So what we're really after isn't a precise rule, but an average trend.

In this example, we're working with a dataset of 35 adult males. Each data point is a pair: a person's weight in kilograms and their height in centimeters. If we plot these on a 2D graph, the points won't fall neatly on a line — and that's okay for real data.

Still, we might want to summarize the overall relationship with a simple mathematical model. And one of the most widely used tools for that is linear approximation. Here, we propose a model of the form y approximately equal to $\theta_0 + \theta_1 x_1$, where y is height and x_1 is weight. The idea is to find a straight line that "best fits" the cloud of points — not perfectly, but in terms of average behavior.

This approach has two huge advantages. First, it's simple and interpretable: we can explain what each parameter means. Second, it gives us a baseline. Once we understand the linear trend, we can check if a more complex model is even necessary.

And to really get a feel for the relationship, we'll visualize the data next. Seeing the points and the fitted line side by side gives us intuition about how well the model captures the data — and where it might fall short.

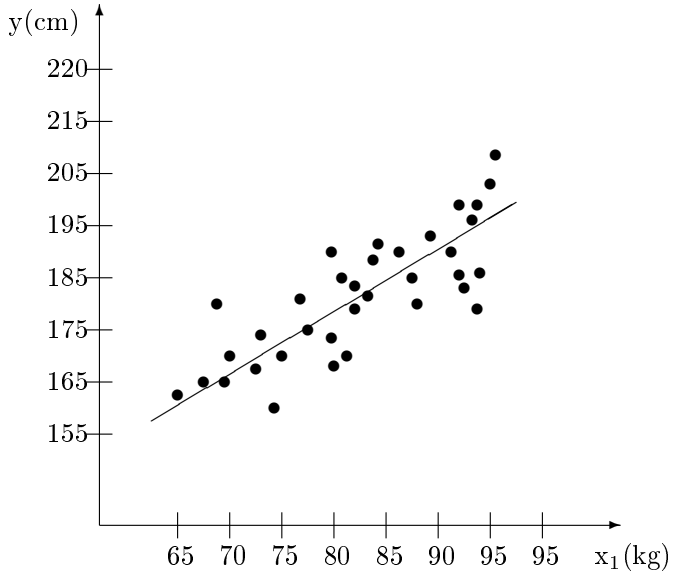

Figure: Data on the ratio of weight and height for 35 men.

Comments

Let's take a look at the actual data.

Here, each dot represents one individual — a pair of weight and height measurements. As you can see, the points are scattered: there's a lot of variation. Some individuals with the same weight have very different heights, and vice versa. This is exactly what we expect from real-world biological data.

But despite the noise, we can still see a clear trend. There's an upward tilt to the cloud of points — as weight increases, height tends to increase as well. This pattern is captured by the straight line we've drawn across the plot. That line is our proposed linear model: it doesn't pass through every point, but it does its best to reflect the average relationship between weight and height.

What's important to notice here is that the model smooths over the individual differences. It gives us a clean, interpretable summary: on average, heavier individuals are taller. Of course, it doesn't explain everything — but that's okay. Our goal at this stage is to capture the overall direction of the relationship, not every detail.

Later, we'll talk about how to actually compute that best-fitting line — and how to measure how well it performs.

Estimating Parameters of the Regression Model

Goal: Determine the parameter values $\theta_1, \ldots, \theta_m$ that make the regression model best describe the observed data.

- After choosing the model form $y \approx \eta(x_1, \ldots, x_m)$, the next task is to find the unknown parameters θ_i.
- A criterion of fit must be introduced to evaluate how well the model matches the experimental data.
- We compare measured values y_j with predicted values $\widetilde{y}_j$.

Deviation of the Linear Model

$$\epsilon_j = y_j - \widetilde{y}_j, \quad j = 1, \ldots, N, \quad \text{where} \quad \widetilde{y}_j = \sum_{i=1}^{m} x_{ji}\theta_i.$$

$$\text{Or equivalently:} \quad y_j = \sum_{i=1}^{m} x_{ji}\theta_i + \epsilon_j$$

$$\text{Matrix form:} \quad y = X\theta + \epsilon,$$

$\theta = (\theta_1, \ldots, \theta_m)^T$ is the vector of unknown parameters,
$\epsilon = (\epsilon_1, \ldots, \epsilon_N)^T$, is the vector of deviations.

Comments

After deciding the model structure, we need to find parameter values that make the model describe the experimental data as well as possible.

To do this, we introduce the concept of error — the difference between the observed value and the predicted value by the model. This error shows how well the model works for each individual observation.

For linear model the predicted value for each data point is calculated as a weighted sum of the input variables, where the weights are the model parameters we want to find.

We can rewrite the relationship by saying that the observed value equals the predicted part plus the error. This highlights that real data usually do not fit perfectly to the model because deviations always exist.

Finally, we represent the whole model compactly using matrix notation: the vector of observed values equals the product of the input data matrix and the parameter vector, plus the vector of errors. This form is very useful for calculations and theoretical analysis.

Criteria for Optimal Parameter Estimation

Error vector and parameter dependence

- The error vector $\epsilon = y - X\theta$ depends on the parameter vector θ, assuming X is fixed.
- The quality of the regression model is thus determined by the choice of θ.

Common optimization criteria

To find the "best" parameters, different criteria can be used:

- Minimize the maximum absolute error: $\max|\epsilon_j|$
- Minimize the sum of absolute errors: $\sum|\epsilon_j|$
- **Minimize the sum of squared errors:** $\epsilon^T\epsilon$

Why least squares is preferred

The last criterion, minimizing the sum of squared errors, is the most widely used due to its computational simplicity and optimality properties.

Comments

With the error vector defined, we now need a meaningful way to choose the best parameter values — that is, those that make the model fit the data as closely as possible.

There are several criteria for this. One option is to minimize the largest absolute error among all predictions. This ensures that no individual prediction is too far off, but it ignores the overall pattern. Another option is to minimize the total sum of absolute deviations. This is more balanced, but the resulting optimization problem is harder to solve analytically.

The most commonly used and most practical criterion is to minimize the sum of squared deviations. That is, we square each residual, then sum them all up. This criterion — known as the least squares method — is not only computationally convenient but also leads to estimators with strong theoretical guarantees, such as unbiasedness and efficiency under classical assumptions.

Its popularity comes precisely from this combination of simplicity and optimality. In most applications, minimizing the squared error is the default strategy, unless there's a specific reason to choose a different loss function.

Definition: Least Squares Estimator

The vector

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^m} (\mathrm{Y} - \mathrm{X}\theta)^{\mathrm{T}} (\mathrm{Y} - \mathrm{X}\theta) = \arg\min_{\theta \in \mathbb{R}^m} \epsilon^{\mathrm{T}} \epsilon$$

is called the empirical least squares estimator (LSE) .

Lemma 1 (Normal Equations)

For any matrix X and vector Y of compatible dimensions, the system

$$\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta = \mathrm{X}^{\mathrm{T}}\mathrm{Y}$$

— called the *system of normal equations* — always has at least one solution. Any vector θ^* satisfying this system is a least squares estimator.

Features of the system of normal equations

- The least squares criterion is quadratic in θ, leading to a convex optimization problem.
- The normal equations provide a closed-form condition for optimality.

Comments

So, as a rule to determine the best-fit parameters for a linear model, we apply the least squares method, minimizing the sum of squared residuals. The parameter vector that achieves this minimum is called the empirical least squares estimator.

The following lemma holds: for any matrix X and any compatible vector Y, the equation system $\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta = \mathrm{X}^{\mathrm{T}}\mathrm{Y}$, known as the system of normal equations, always has at least one solution. Any vector that satisfies these equations is a least squares estimator.

This formulation is critical in both theoretical and computational aspects of regression analysis. It allows us to find an optimal estimate analytically — at least when the normal matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is invertible. Even if it's not, a solution still exists, possibly non-unique.

The optimization problem itself is convex, since the objective function is quadratic in θ. This means that any solution to the normal equations corresponds to a global minimum. The name "normal equations" comes from classical least squares theory and is standard terminology in linear regression.

Lemma 1: Proof (part 1)

Proof: To begin with, we show that there is always a solution to a system of normal equations. For any matrix A, denote by $\mathcal{L}(A)$ the linear span of its columns.
We claim that:

$$\mathcal{L}(A^T) = \mathcal{L}(A^T A).$$

- Let b be a vector orthogonal to $\mathcal{L}(A^T)$. Then

$$b^T A^T = 0 \quad \Rightarrow \quad b^T A^T A = 0,$$

so $b \perp \mathcal{L}(A^T A)$.

- Conversely, if $b \perp \mathcal{L}(A^T A)$, then

$$b^T A^T A = 0 \quad \Rightarrow \quad (Ab)^T Ab = 0 \quad \Rightarrow \quad Ab = 0 \quad \Rightarrow \quad b^T A^T = 0.$$

So $b \perp \mathcal{L}(A^T)$.

Hence, the spaces coincide:

$$\mathcal{L}(A^T) = \mathcal{L}(A^T A).$$

Thus, for any $X^T Y \in \mathcal{L}(X^T)$, there exists θ^* such that

$$X^T X\theta^* = X^T Y.$$

Comments

We begin the proof by showing that the system of normal equations always has a solution.

To do this, we prove that for any matrix A, the space spanned by the rows of A, which is the same as the column space of A transposed, coincides with the column space of the matrix $A^T A$.

To show this, take any vector b orthogonal to the column space of A transposed. Then $b^T A^T = 0$, and multiplying by A again yields zero. This implies b is orthogonal to the column space of $A^T A$.

Conversely, if b is orthogonal to the column space of A transposed A, then the product $(Ab)^T Ab = 0$, so Ab is zero, and hence $b^T A^T$ is also zero.This proves the spaces coincide. It follows that the vector X transposed Y lies in the column space of $X^T X$, so the equation $X^T X\theta = X^T Y$ has a solution θ^*.

Lemma 1: Proof (part 2)

Let $\widehat{\theta}$ be any solution to the normal equations:

$$\mathrm{X}^{\mathrm{T}}\mathrm{X}\widehat{\theta} = \mathrm{X}^{\mathrm{T}}\mathrm{Y}.$$

Then for arbitrary θ:

$$\begin{aligned}&(\mathrm{Y} - \mathrm{X}\theta)^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\theta) = \\ &= (\mathrm{Y} - \mathrm{X}\widehat{\theta} + \mathrm{X}(\widehat{\theta} - \theta))^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\widehat{\theta} + \mathrm{X}(\widehat{\theta} - \theta)) = \\ &= (\mathrm{Y} - \mathrm{X}\widehat{\theta})^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\widehat{\theta}) + (\widehat{\theta} - \theta)^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{X}(\widehat{\theta} - \theta) \geq \\ &\geq (\mathrm{Y} - \mathrm{X}\widehat{\theta})^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\widehat{\theta}).\end{aligned}$$

The cross term vanishes due to:

$$(\widehat{\theta} - \theta)^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\widehat{\theta}) = 0.$$

Therefore, $\widehat{\theta}$ minimizes the squared error. The Lemma is proved. □

Remark

If $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is nonsingular, then the normal system has a unique solution:

$$\widehat{\theta} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}.$$

Comments

Now we show that any solution to a system of normal equations is least squares estimate. To be precise let θ hat be any solution to the system of normal equations. Then we compute the squared error norm of any arbitrary θ by rewriting the expression in terms of θ hat. This is done using a standard identity from linear algebra.

The key point is that the cross-term vanishes due to the fact that θ hat solves the normal equations. As a result, the squared error of any θ equals the squared error of θ hat plus a non-negative term. This implies that the minimum is achieved exactly when $\theta = \widehat{\theta}$. Therefore, any solution to the normal equations minimizes the residual norm and is indeed least squares estimate. The Lemma is proved.

We also note that if the matrix X transposed times X is nonsingular, then the normal equations have a unique solution. This unique estimator is then given explicitly as $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$.

Mathematical formulation

The classical linear regression model is written as:

$$y_j = \sum_{i=1}^{m} x_{ji}\theta_i + \epsilon_j, \quad j = 1, \dots, N, \tag{1}$$

or in matrix form:

$$Y = X\theta + \epsilon. \tag{2}$$

- $\theta = (\theta_1, \dots, \theta_m)^T$ — vector of unknown coefficients.
- $Y = (y_1, \dots, y_N)^T$ — vector of observed responses.
- $X = (x_{ij})_{i=1,j=1}^{m\ N}$ — design matrix:

$$X = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{Nm} \end{pmatrix}$$

- $\epsilon = (\epsilon_1, \dots, \epsilon_N)^T$ — vector of random errors.

Comments

In the regression setting, the observed output is modeled as a deterministic part plus an error term. The deterministic part is a linear combination of the inputs, with unknown coefficients that describe how each input influences the output. The error term accounts for noise and other unmodeled effects.

The classical linear regression model assumes that each observation y_j is equal to the sum of the input values x_{ji} multiplied by unknown coefficients θ_i, plus a random error ϵ_j. This can be written as: y_j equals the sum over i from 1 to m of x_{ji} times θ_i, plus ϵ_j, or j from 1 to N.

To make the notation more compact and the analysis more convenient, the model is expressed in matrix form as: Y equals X times θ plus ϵ, where Y is the column vector of all observed outputs, X is the design matrix that stores all input values, θ is the vector of unknown coefficients, and ϵ is the vector of random errors.

Here, the vector ϵ is modeled as a random variable — typically assumed to follow a normal distribution centered at zero. This stochastic interpretation reflects the fact that measurements are never perfectly accurate and that real-world systems include inherent randomness.

By explicitly including this randomness in the model, statistical inference becomes possible: one can estimate the coefficients, assess the quality of the model, and make probabilistic predictions.

Classical Linear Regression: Notation and Assumptions

Model Notation: We denote the classical linear regression model by the triplet

$$(\mathrm{Y}, \mathrm{X}\theta, \Sigma),$$

where Σ is a fixed $\mathrm{N} \times \mathrm{N}$ covariance matrix of the random errors:

$$\Sigma = \mathrm{E}[\epsilon\epsilon^{\mathrm{T}}] = \|\mathrm{E}[\epsilon_{\mathrm{i}}\epsilon_{\mathrm{j}}]\|_{\mathrm{i,j}=1}^{\mathrm{N}}.$$

(1) **Unbiasedness:** $\mathrm{E}[\epsilon_{\mathrm{i}}] = 0$.
(2) **Homoscedasticity:** $\mathrm{E}[\epsilon_{\mathrm{i}}^2] = \sigma^2$.
(3) **Uncorrelated Errors:** $\mathrm{E}[\epsilon_{\mathrm{i}}\epsilon_{\mathrm{j}}] = 0$ for $\mathrm{i} \neq \mathrm{j}$.

(a) **Estimator Unbiasedness:** $\mathrm{E}[\hat{\theta}] = \theta$.
(b) **Minimum Variance:** For any vector z of appropriate dimension and any unbiased estimator $\tilde{\theta}$, the inequality

$$\mathrm{D}\left(\mathrm{z}^{\mathrm{T}}(\hat{\theta} - \theta)\right) \leq \mathrm{D}\left(\mathrm{z}^{\mathrm{T}}(\tilde{\theta} - \theta)\right)$$

holds, where D denotes the covariance matrix.
(c) **Linearity:** $\hat{\theta} = \mathrm{SY}$, where S is a matrix independent of Y.

Comments

From now on, to represent the classical regression model, we will use the following triplet: the observed response vector, the deterministic part involving the design matrix and parameters, and the covariance matrix of the error terms. This covariance matrix is fixed and encodes how errors vary and how they might be correlated.

As a rule, the following standard assumptions about errors and parameter estimates are accepted.

First, we assume errors have zero mean, which means that on average the errors do not systematically bias the observations. This ensures that our model predictions are centered around the true values rather than consistently over- or underestimating them.

Second, the errors are assumed to have equal variance, called homoscedasticity. This reflects the idea that the variability of errors is uniform across all observations, so no particular measurement is inherently more uncertain than ‘another. This assumption simplifies both estimation and inference.

Third, we assume errors are uncorrelated, meaning the error in one observation does not influence the error in another. This is crucial because correlations would imply some hidden structure or dependence in the noise, which requires more complex modeling.

Regarding parameter estimates, the goal is to find estimators that, on average, correctly recover the true parameters (unbiasedness). Among all unbiased estimators, we prefer those with minimal variance for any linear combination of parameters, ensuring the estimates are as precise as possible.

Finally, insisting that the estimator be a linear function of the observed data allows for elegant mathematical treatment and computational efficiency. This linearity condition means the estimator can be expressed as a fixed matrix multiplying the data vector, independent of the data realization itself.

Variance Minimization Condition and BLUE

Let's comment on condition (b).

- Let $\mathrm{D}_{\tilde{\theta}} = \mathrm{E}[(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^{\mathrm{T}}]$ denote the covariance matrix of an unbiased estimator $\tilde{\theta}$.
- For any vector z of compatible dimension,

$$\mathrm{D}\big(\mathrm{z}^{\mathrm{T}}(\tilde{\theta} - \theta)\big) = \mathrm{E}\big(\mathrm{z}^{\mathrm{T}}(\tilde{\theta} - \theta)\big)^2 = \mathrm{z}^{\mathrm{T}}\mathrm{D}_{\tilde{\theta}}\,\mathrm{z},$$

  where the last equality holds due to the unbiasedness of $\tilde{\theta}$.
- The variance minimization condition (b) implies:

$$\mathrm{z}^{\mathrm{T}}\mathrm{D}_{\hat{\theta}}\,\mathrm{z} \le \mathrm{z}^{\mathrm{T}}\mathrm{D}_{\tilde{\theta}}\,\mathrm{z},$$

  i.e., the matrix $\mathrm{D}_{\tilde{\theta}} - \mathrm{D}_{\hat{\theta}}$ is positive semi-definite.
- Note: An estimator satisfying this condition does not always exist.

Definition: Best Linear Unbiased Estimator (BLUE)

An estimator that satisfies conditions (a) unbiasedness, (b) minimal variance, and (c) linearity is called the Best Linear Unbiased Estimator (BLUE).

Comments

Let's now take a closer look at condition (b), the variance minimization requirement. At first glance, it may seem unclear why we're interested not simply in minimizing the variances of the individual components of the estimator vector $\hat{\theta}$, but instead in minimizing the variance of all possible linear combinations $\mathrm{z}^{\mathrm{T}}\hat{\theta}$, where z is any fixed vector.

Here's why. In many practical problems, we're not always interested in all components of the parameter vector individually. Sometimes we care about a specific function of the parameters — say, a contrast like θ_1 minus θ_2, or a predicted value at some point, which also ends up being a linear combination of parameters. Therefore, it is natural to assess the precision of $\hat{\theta}$ not just globally, but in every possible direction in parameter space — i.e., for all vectors z.

Condition (b) says: if we take any linear combination $\mathrm{z}^{\mathrm{T}}\hat{\theta}$, then its variance should be no greater than for any other unbiased estimator $\tilde{\theta}$. This guarantees that $\hat{\theta}$ is not just good on average, but it delivers the tightest possible confidence intervals for any estimable quantity derived from θ.

In matrix terms, this condition means that the difference between the covariance matrices of $\tilde{\theta}$ and $\hat{\theta}$ is a positive semi-definite matrix. That is: for all z, the quadratic form with this difference is non-negative.

An estimator satisfying conditions (a), (b), and (c) is called a BLUE — Best Linear Unbiased Estimator. It is the benchmark against which all other linear estimators are judged.

Example: Linear Regression Model

Consider a linear regression model of the form:

$$\eta(t) = a + bt.$$

This model describes, for instance, the change in length of a metal rod as a function of temperature.
Measurements are taken at points $t_1, \ldots, t_N$:

$$y_j = a + bt_j + \epsilon_j, \quad j = 1, \ldots, N.$$

We represent the model in matrix form:

$$Y = X\theta + \epsilon, \quad \theta = \begin{pmatrix} a \\ b \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix},$$

$$X = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_N \end{pmatrix}, \quad X\theta = \begin{pmatrix} a + bt_1 \\ \vdots \\ a + bt_N \end{pmatrix}.$$

Comments

Let's consider a simple example. Here we're modeling how a metal rod's length changes with temperature - a classic physics experiment. We assume this relationship is linear and can be described by an equation where the outcome depends on the sum of two components: a constant parameter plus another parameter multiplied by temperature. Both parameters are unknown and need to be estimated.

We take measurements at different temperature points - from the first to the nth observation - obtaining corresponding measured values. Each measurement differs from the true value by some random error. Thus, each observed value equals the sum of: (1) the constant parameter, (2) temperature multiplied by the second parameter, and (3) random measurement error.

This model can be conveniently expressed in matrix form. Here, the parameter vector contains two elements: the intercept (constant term) and the temperature coefficient. The design matrix has two columns: one column of ones and another column of temperature values at each measurement. The observation vector contains all measured lengths, while the error vector collects all random deviations.

The matrix formulation offers key advantages: it lets us use standard linear algebra tools like pseudoinverse for parameter estimation, analyze estimator properties, construct confidence intervals, and test hypotheses. Moreover, it naturally extends to more complex models and simplifies computations with large datasets.

Definition

For a classical linear regression model $(Y, X\theta, \Sigma)$ with the nonsingular matrix X^TX, the vector

$$\widehat{\theta} = (X^TX)^{-1}X^TY$$

is called the *Ordinary Least Squares estimator (OLS estimator)*.

Example: Linear Model

Consider again the model describing the dependence of a metal rod's length on temperature:

$$y_j = a + bt_j + \epsilon_j, \quad j = 1, \ldots, N,$$

where $t_1, \ldots, t_N$ are given real numbers.
The matrix X^TX takes the form:

$$X^TX = \begin{pmatrix} 1 & \ldots & 1 \\ t_1 1 & \ldots & t_N \end{pmatrix} \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_N \end{pmatrix} = \begin{pmatrix} N & \sum t_j \\ \sum t_j & \sum t_j^2 \end{pmatrix}.$$

Comments

Earlier, we introduced the empirical least squares estimate. For the mathematical model of linear regression represented by the triple $(Y, X\theta, \Sigma)$, we now define the standard Ordinary Least Squares estimator as: the inverse of X^TX, multiplied by X^T, and then by the vector Y. Note that in this case, the estimator is obtained as the solution to the system of normal equations. It is assumed here that the matrix X^TX is nonsingular. Later on, we will generalize this definition to the case where this matrix is singular.

Let us return to the example we considered earlier — the model describing the dependence of a metal rod's length on temperature. In this case, the matrix X^TX is nonsingular and has the following form: the number of observations N, the sum of t_j, the sum of t_j again, and the sum of squared t_j.

Example (continued)

Using the standard formula for the inverse of a 2×2 matrix, we obtain:

$$(X^T X)^{-1} = \begin{pmatrix} \frac{\sum t_j^2}{\Delta} & -\frac{\sum t_j}{\Delta} \\ -\frac{\sum t_j}{\Delta} & \frac{N}{\Delta} \end{pmatrix}, \quad \Delta = N \sum t_j^2 - \left(\sum t_j\right)^2.$$

$$X^T Y = \begin{pmatrix} \sum y_j \\ \sum y_j t_j \end{pmatrix}.$$

Hence, the OLS estimator is:

$$\begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix} = (X^T X)^{-1} X^T Y = \begin{pmatrix} \frac{\sum t_j^2 \sum y_j - \sum t_j \sum y_j t_j}{\Delta} \\ \frac{N \sum y_j t_j - \sum t_j \sum y_j}{\Delta} \end{pmatrix}.$$

- This is the explicit form of the OLS estimates for the intercept and slope in the simple linear model.
- These formulas depend only on the sample sums and can be computed directly from the data.

Comments

Using the standard formula for the inverse of a two-by-two matrix, we obtain an explicit expression for the inverse of the matrix $X^T X$.

Multiplying the inverse of $X^T X$ with $X^T Y$, we obtain explicit formulas for the estimated intercept and slope — that is, the least squares estimates of the coefficients a and b in our linear model. These estimates depend only on basic sample statistics and are easy to compute in practice. They form the foundation of many methods used in applied regression analysis.

Theorem 1 (Gauss–Markov)

Consider the classical linear regression model $(Y, X\theta, \sigma^2 I_N)$, where:

- σ^2 is the common variance of the errors.
- I_N is the identity matrix of size N,
- the error vector ϵ satisfies the assumptions (1)–(3):
  (1) $E[\epsilon] = 0$,
  (2) $\mathrm{Cov}(\epsilon) = \sigma^2 I_N$,
  (3) the components of ϵ are uncorrelated.
- and the matrix $X^T X$ is nonsingular.

Then the vector

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

is the *Best Linear Unbiased Estimator (BLUE)* of θ. In other words, it has the *minimum variance* among all linear unbiased estimators.

Its covariance matrix is given by:

$$D_{\widehat{\theta}} = \sigma^2 (X^T X)^{-1}.$$

Interpretation

Comments

On this slide, we present the Gauss–Markov theorem, which is a key result in the theory of linear regression. We consider the classical model, in which the error vector has zero mean, uncorrelated components, and equal variance — in other words, the covariance matrix of the errors is σ^2 times the identity matrix. In addition, we assume that the matrix $X^T X$ is invertible.

Under these conditions, the Ordinary Least Squares estimator, which is $(X^T X)^{-1} X^T Y$, turns out to be the best possible — that is, it is the most efficient among all linear and unbiased estimators. This is what we mean when we say that it is the best linear unbiased estimator, or BLUE.

The theorem also gives us the exact formula for the covariance matrix of the estimator: it equals σ^2 times the inverse of $X^T X$. This matrix reflects the precision of our estimates and plays a central role in evaluating statistical reliability.

In summary, this theorem not only guarantees the optimality of the estimator under the given assumptions, but also allows us to quantify the variability of the estimated parameters.

Lemma 2

The OLS estimator is a linear and unbiased estimator, i.e., it satisfies conditions (a) and (c):

(a) Estimator Unbiasedness: $\mathrm{E}[\hat{\theta}] = \theta$.

(c) Linearity: $\hat{\theta} = \mathrm{SY}$, where S is a matrix independent of Y.

Proof: Linearity of the OLS estimator follows directly from its form: the matrix $\mathrm{S} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$.
To verify unbiasedness, consider the expectation: since $\mathrm{E}(\mathrm{Y}) = \mathrm{X}\theta$, we get $\mathrm{E}(\widehat{\theta}) = \mathrm{E}(\mathrm{SY}) = \mathrm{S} \cdot \mathrm{E}(\mathrm{Y}) = \mathrm{SX}\theta = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta = \theta$. Lemma is proved. □

Lemma 3

A linear estimator $\widetilde{\theta} = \mathrm{AY}$ is unbiased if and only if $\mathrm{AX} = \mathrm{I}$.

Comment

Lemma 3 is the key criterion for verifying the unbiasedness of any linear estimator. It will be used in the proof of the Gauss–Markov theorem.

Comments

This slide presents two supporting lemmas that we need in order to prove the Gauss–Markov theorem.

The first lemma states that the OLS estimator is both linear and unbiased. Linearity means that the estimator is equal to a fixed matrix times the observation vector Y. In this case, the matrix is defined as the inverse of $\mathrm{X}^{\mathrm{T}}\mathrm{X}$, multiplied by X^{T}. This property is important because it means the estimator reacts to data in a predictable, linear way.

To verify that the estimator is unbiased, we consider the expected value of the estimator. Due to the fact that the expected value of error is zero, we have that the expected value of Y is equal to X times θ. When we substitute this into the formula for the estimator, all the matrices cancel out, and we obtain exactly θ. This confirms that the estimator is unbiased.

The second lemma gives us a general condition for checking whether a linear estimator is unbiased. Suppose we define an estimator as a matrix A times the vector Y. Then this estimator is unbiased if and only if the product of A and the matrix X is equal to the identity matrix.

This second lemma will serve as the main technical tool in the proof of the Gauss–Markov theorem, which we will see shortly.

Proof of Lemma 3

Proof:

Sufficiency: Assume $AX = I$. Then:

- The estimator has the form: $\widetilde{\theta} = AY$.
- Since $E(Y) = X\theta$, we have:

$$E[\widetilde{\theta}] = AE[Y] = AX\theta = \theta.$$

- Hence, $\widetilde{\theta}$ is unbiased.

Necessity: Suppose $\widetilde{\theta} = AY$ is unbiased.

- Then $E[\widetilde{\theta}] = AX\theta = \theta$ for all $\theta \in \mathbb{R}^m$.
- This implies that:

$$AX\theta = \theta \quad \text{for all } \theta.$$

- Let us test this identity on the standard basis vectors:

$$\theta = e_i = (0, \dots, 0, 1, 0, \dots, 0)^T \text{ with one in the i-th place..}$$

- Then $AXe_i = e_i$, so the i-th column of AX equals the i-th column of the identity matrix.
- Repeating this for all $i = 1, \dots, m$, we conclude:

$$AX = I.$$

The Lemma is proved. □

Comments

This slide presents a full proof of Lemma 3, which defines the necessary and sufficient conditions for the estimator to be unbiased.

We begin with sufficiency. Suppose that the matrix A multiplied by X equals the identity. Then the estimator AY has expectation equal to A times the expected value of Y. Since the expectation of Y is X times θ, we get $AX\theta$, which is simply θ. Therefore, the estimator is unbiased.

For necessity, we start from the assumption that AY is an unbiased estimator. This means that the expectation of AY must be equal to θ for all possible values of θ. In other words, $AX\theta$ equals θ for any vector θ.

To show that this leads to AX being the identity matrix, we substitute standard basis vectors one by one into the expression. Each substitution tells us that one column of AX must match the identity matrix. After going through all components, we conclude that AX equals the identity.

This completes the proof.

Lemma 4

Under the assumptions of the Gauss–Markov Theorem, the covariance matrix of any linear unbiased estimator $\widetilde{\theta} = \mathrm{AY}$ has the form:

$$\mathrm{D}_{\widetilde{\theta}} = \sigma^2 \mathrm{A}\mathrm{A}^{\mathrm{T}}.$$

In particular, for the OLS estimator:

$$\mathrm{D}_{\widehat{\theta}} = \sigma^2 (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}.$$

Proof:

Consider the definition of the covariance matrix:

$$\mathrm{D}_{\widetilde{\theta}} = \mathrm{E}[(\widetilde{\theta} - \mathrm{E}[\widetilde{\theta}])(\widetilde{\theta} - \mathrm{E}[\widetilde{\theta}])^{\mathrm{T}}].$$

Since $\widetilde{\theta} = \mathrm{AY}$ and $\mathrm{E}[\widetilde{\theta}] = \theta$, we have:

$$\mathrm{D}_{\widetilde{\theta}} = \mathrm{E}[(\mathrm{AY} - \theta)(\mathrm{AY} - \theta)^{\mathrm{T}}].$$

Comments

This lemma establishes the form of the covariance matrix of a linear, unbiased estimator under the same assumptions as in the Gauss–Markov theorem.

Since the estimator is linear, it can be written as a matrix A multiplied by the vector of observations Y. Furthermore, the estimator is unbiased, which means that the expected value of A times Y equals the true parameter vector θ.

To find the covariance matrix of this estimator, we recall the general formula: it's the expected value of the deviation from the mean, multiplied by its own transpose.

In our case, the deviation is A times Y minus θ. So the covariance matrix becomes the expected value of that expression times its transpose.

Variance of Linear Unbiased Estimators (continued)

Since $\mathrm{E}[\mathrm{AY}] = \mathrm{E}[\mathrm{Y}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}]^{\mathrm{T}} = \theta$ we have:

$$\mathrm{E}[(\mathrm{AY}-\theta)(\mathrm{AY}-\theta)^{\mathrm{T}}] = \mathrm{E}[\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}] - \theta\theta^{\mathrm{T}}.$$

To compute $\mathrm{E}[\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}]$, substitute the expression for Y:

$$\mathrm{Y} = \mathrm{X}\theta + \varepsilon \quad \Rightarrow \quad \mathrm{YY}^{\mathrm{T}} = (\mathrm{X}\theta+\varepsilon)(\mathrm{X}\theta+\varepsilon)^{\mathrm{T}}.$$

Then:

$$\mathrm{E}[\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}] = \mathrm{AE}[(\mathrm{X}\theta+\varepsilon)(\mathrm{X}\theta+\varepsilon)^{\mathrm{T}}]\mathrm{A}^{\mathrm{T}}.$$

Using independence and zero mean of ε, we get:

$$\mathrm{E}[\mathrm{YY}^{\mathrm{T}}] = \mathrm{X}\theta\theta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}} + \sigma^2\mathrm{I} \quad \Rightarrow \quad \mathrm{E}[\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}] = \mathrm{AX}\theta\theta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}} + \sigma^2\mathrm{AA}^{\mathrm{T}}.$$

Since $\mathrm{AX} = \mathrm{I}$, this simplifies to:

$$\mathrm{E}[\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}] = \theta\theta^{\mathrm{T}} + \sigma^2\mathrm{AA}^{\mathrm{T}}.$$

Subtracting $\theta\theta^{\mathrm{T}}$, we conclude:

$$\mathrm{D}_{\widetilde{\theta}} = \sigma^2\mathrm{AA}^{\mathrm{T}}.$$

For the OLS estimator, where $\mathrm{A} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$, we have:

$$\mathrm{D}_{\widehat{\theta}} = \sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} = \sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}. \qquad \square$$

Comments

To complete the proof, note that since the expected value of AY equals θ we can rewrite the covariance matrix of the estimator as expectation of $\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}$ minus $\theta\theta^{\mathrm{T}}$.

To compute this, we use the fact that the vector of observations Y can be written as X times θ plus ϵ, where ϵ is the vector of random errors. Then the matrix Y times Y^{T} becomes the product of X times θ plus ϵ with its transpose. Taking expectation of the resulting expression, and using linearity along with the assumption that the error vector ϵ has zero mean and is uncorrelated with the regressors, we obtain: expectation of Y times Y^{T} equals X times θ times θ^{T} times X^{T} plus σ^2 times the identity matrix.

Multiplying by A from the left and by A^{T} from the right, we get that the desired expectation is equal to A times X times θ times θ^{T} times X^{T} times A^{T} plus σ^2 times A times A^{T}.

Now, since we assumed that A times X equals the identity matrix, this simplifies to θ times θ^{T} plus σ^2 times A times A^{T}.

Subtracting the outer product θ times θ^{T}, we arrive at the final result: the covariance matrix of the estimator equals σ^2 times A times A^{T}.

Finally, in the case of the OLS estimator, the matrix A is given by the inverse of $\mathrm{X}^{\mathrm{T}}\mathrm{X}$, all multiplied by X^{T}. Substituting this into the general formula and simplifying, we get that the covariance matrix of the OLS estimator is equal to σ^2 times the inverse of $\mathrm{X}^{\mathrm{T}}\mathrm{X}$, which completes the proof.

Proof of Theorem 1:
Let us verify the matrix inequality:

$$\mathrm{D}_{\widehat{\theta}} \leq \mathrm{D}_{\widetilde{\theta}}.$$

Denote $\mathrm{S} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$, so that $\widehat{\theta} = \mathrm{SY}$. For any linear unbiased estimator $\widetilde{\theta} = \mathrm{AY}$, we compute:

$$\mathrm{D}_{\widetilde{\theta}} = \mathrm{D}(\mathrm{AY}) = \mathrm{D}((\mathrm{A}-\mathrm{S})\mathrm{Y} + \mathrm{SY}).$$

Using the identity $\mathrm{D}(\mathrm{U}+\mathrm{V}) = \mathrm{D}(\mathrm{U}) + \mathrm{D}(\mathrm{V})$ when U and V are uncorrelated, we obtain:

$$\mathrm{D}_{\widetilde{\theta}} = \mathrm{D}((\mathrm{A}-\mathrm{S})\mathrm{Y}) + \mathrm{D}(\mathrm{SY}) = \mathrm{D}((\mathrm{A}-\mathrm{S})\mathrm{Y}) + \mathrm{D}_{\widehat{\theta}} \geq \mathrm{D}_{\widehat{\theta}}.$$

The cross term vanishes because $(\mathrm{A}-\mathrm{S})\mathrm{X} = \mathrm{I} - \mathrm{I} = 0$, and thus:

$$\mathrm{E}[(\mathrm{A}-\mathrm{S})\mathrm{YY}^{\mathrm{T}}\mathrm{S}^{\mathrm{T}}] = (\mathrm{A}-\mathrm{S})\mathrm{E}[\mathrm{YY}^{\mathrm{T}}]\mathrm{S}^{\mathrm{T}} = (\mathrm{A}-\mathrm{S})(\mathrm{X}\theta\theta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}} + \sigma^2\mathrm{I})\mathrm{S}^{\mathrm{T}} = 0.$$

This concludes the proof. □

Comments

Now we return to the proof of the Gauss–Markov Theorem.Our goal is to show that the covariance matrix of the ordinary least squares estimator is less than or equal to the covariance matrix of any other linear unbiased estimator — in the sense of matrix inequality. This means that the OLS estimator is the best, in the sense of having minimal variance, within the class of all linear unbiased estimators.

We denote the OLS matrix by capital S, where S equals the inverse of $\mathrm{X}^{\mathrm{T}}\mathrm{X}$, multiplied by X^{T}. Then the OLS estimator, $\hat{\theta}$, is equal to S times Y.

Now consider any linear unbiased estimator, denoted by $\tilde{\theta}$, which has the form A times Y. We rewrite this as: A times Y equals the sum of two parts — namely, A minus S times Y, plus S times Y.

Then, the covariance matrix of $\tilde{\theta}$ equals the covariance of the first term, A minus S times Y, plus the covariance of the second term, S times Y.

We are allowed to add the covariances because the two components are uncorrelated. This is due to the fact that the matrix A minus S multiplied by X gives zero. Therefore, A minus S times Y is uncorrelated with S times Y, and the mixed covariance term vanishes.

As a result, the covariance matrix of any linear unbiased estimator is always greater than or equal to that of the OLS estimator. This proves that the OLS estimator has the smallest possible variance among all linear unbiased estimators. The theorem is thus proven.

Example: Exact Data and Quadratic Approximation

Example

$$y = \frac{2}{4 - 3x}$$

This true function is known, but used only for analyzing model performance — not for building the model.

Measurement Points:

- $x_1 = 0, \quad y_1 = 1/2$
- $x_2 = \frac{2}{3}, \quad y_2 = 1$
- $x_3 = 1, \quad y_3 = 2$

Model: Quadratic Regression (No Measurement Error)

$$\eta(x, \theta) = \theta_0 + \theta_1 x + \theta_2 x^2$$

Vectors and Matrix:

$$Y = \begin{pmatrix} 1/2 \\ 1 \\ 2 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & \frac{2}{3} & \frac{4}{9} \\ 1 & 1 & 1 \end{pmatrix}$$

Comments

In this example, we consider a situation where the true dependence between input and output is exactly known and given by the function: y equals two divided by the quantity four minus three x. This information is not used to construct the model — it is included solely to evaluate its accuracy.

We conduct measurements at three input points: x equals zero, x equals two-thirds, and x equals one. Evaluating the true function at these points, we obtain the corresponding outputs: one-half, one, and two. These values form the observation vector Y.

The model we use for approximation is a quadratic regression of the form: θ_0 plus θ_1 times x plus θ_2 times x squared. This means we assume that the data can be well-approximated by a second-degree polynomial.

Additionally, we assume that there are no measurement errors. Of course, this is an artificial assumption made only to simplify the analysis.

The design matrix X is constructed by applying the basis functions — constant, linear, and quadratic — to each of the three input values. At x equals zero, the row is: one, zero, zero. At x equals two-thirds, the row is: one, two-thirds, and four-ninths. At x equals one, the row is: one, one, one. These will be used to compute the least squares estimate and analyze how well the model fits the data.

Example (continued)

To compute the OLS estimator, we start by finding the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$:

$$\mathrm{X}^{\mathrm{T}}\mathrm{X} = \begin{pmatrix} 3 & \frac{5}{3} & \frac{13}{9} \\ \frac{5}{3} & \frac{13}{9} & \frac{35}{27} \\ \frac{13}{9} & \frac{35}{27} & \frac{97}{81} \end{pmatrix}, \quad (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} = \begin{pmatrix} 1 & -\frac{5}{2} & \frac{3}{2} \\ -\frac{5}{2} & \frac{61}{2} & -30 \\ \frac{3}{2} & -30 & \frac{63}{2} \end{pmatrix}.$$

Then, the least squares estimator is:

$$\widehat{\theta} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y} = \begin{pmatrix} 1/2 \\ -3/4 \\ 9/4 \end{pmatrix}.$$

Thus, the original function $\frac{2}{4-3x}$ is approximated by a quadratic function obtained via the method of least squares.

Note: This quadratic approximation matches the Lagrange interpolating polynomial for the same nodes.

Comments

Now let's compute the least squares estimates step by step. First, we multiply the design matrix X by its transpose - this gives us the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ shown here. Then we calculate the inverse matrix.

Finally we multiply this inverse by $\mathrm{X}^{\mathrm{T}}\mathrm{Y}$ (which combines our observations with the basis functions), we get our parameter estimates: θ_0 equals one-half, θ_1 equals minus three-fourths, and θ_2 equals nine-fourths.

These coefficients define our quadratic approximation. Interestingly, this exact same polynomial would emerge if we used Lagrange interpolation - but here we derived it through least squares, which is more general. The perfect fit at our three points was guaranteed because we have three parameters and three exact observations - but remember, this is a special case with no measurement errors.

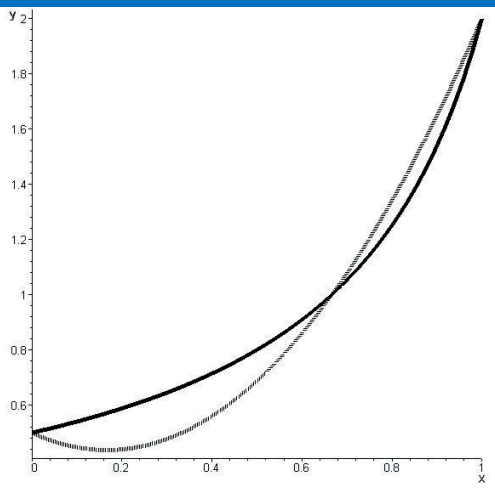

Figure: Approximation of the function $y = \frac{2}{4-3x}$ (solid line) by the function $\bar{y} = 1/2 - 3/4x + 9/4x^2$ (dashed line) constructed using the least squares method.

Comments

This figure shows the result of approximating the function y equals two divided by the quantity four minus three x using the least squares method. The original nonlinear function is drawn with a solid curve, while the dashed curve represents the quadratic approximation defined by the formula: $\bar{y}$ equals one half minus three fourths x plus nine fourths x squared.

As we have already noted, in this particular case the least squares approximation coincides with the Lagrange interpolating polynomial. However, this is not a general feature of least squares methods. It occurs here only because the number of parameters in the model is equal to the number of observations and there are no observation errors in our measurements - we're using the exact function values. Let me emphasize that in the general case, the least squares method does not interpolate the data. Instead, it finds the function that minimizes the total squared error between the predicted values and the observed ones. That is the essence of least squares approximation — it aims not to match the data exactly, but to provide the best possible overall fit in terms of squared deviations.

A key point to emphasize is that the result of a least squares approximation depends not only on the form of the model but also on the locations of the data points. If we had chosen different x values, even with the same model, we would have obtained different coefficients. Thus, the design of the experiment — that is, the selection of input values at which the system is observed — plays a crucial role in the quality of the approximation.

This leads us naturally to the topic of optimal experimental design, which studies how to choose data points to achieve the most informative or precise estimates. In practical applications, especially when experiments are costly or time-consuming, choosing the right points can significantly improve the accuracy of estimation. We will explore this topic in more detail when we study optimality criteria and planning strategies in regression analysis.

Estimability of Linear Parametric Functions

Definition: OLS

A linear parametric function $\tau = \mathrm{T}\theta$ $(\mathrm{T} \in \mathbb{R}^{k\times m},\ k \in \{1,\dots,m\})$ is called estimable if there exists an unbiased estimator of the form $\widehat{\tau} = \mathrm{A}\mathrm{Y}$ with $\mathrm{A} \in \mathbb{R}^{k\times N}$.

Lemma 5: Equivalent Conditions

Consider the classical linear regression model $\mathrm{Y} = \mathrm{X}\theta + \varepsilon$ under the assumption that $\mathbb{E}[\varepsilon] = 0$. Then the following statements are equivalent:

(a) Each row of T lies in the row space of X, i.e., $\mathcal{L}(\mathrm{T}^{\mathrm{T}}) \subseteq \mathcal{L}(\mathrm{X}^{\mathrm{T}}) = \mathcal{L}(\mathrm{X}^{\mathrm{T}}\mathrm{X})$;

(b) The value of $\mathrm{T}\widehat{\theta}$ is the same for all solutions of the system of normal equations $\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta = \mathrm{X}^{\mathrm{T}}\mathrm{Y}$;

(c) The parametric function $\tau = \mathrm{T}\theta$ is estimable.

Note: Estimability ensures that a parametric function $\tau = \mathrm{T}\theta$ can be recovered from data without bias, using only the information contained in the design matrix X.

Comments

We now generalize the classical estimation problem. Previously, we estimated the full parameter vector θ, but in practice, we are often interested in specific functions of parameters—for example, differences between them or averages. In this context, we define a parametric function τ as T times θ, where T is a known matrix of size k by m. The number k indicates how many different linear combinations of parameters we are consider.

Let us now state the formal definition. We say that a parametric function $\tau = \mathrm{T}\theta$ is estimable if there exists an unbiased linear estimator of the form $\hat{\tau} = \mathrm{A}\mathrm{Y}$, where A is a matrix of size k by N. That is, $\hat{\tau}$ is a linear function of the data, and its expected value equals τ.

This is not just a technical detail. In many statistical applications, we cannot recover the full vector θ, but we can still estimate certain combinations of its components—and this definition tells us what it means to do so without bias.

The lemma on the slide provides necessary and sufficient conditions for estimability in the classical linear model $\mathrm{Y} = \mathrm{X}\theta + \epsilon$, assuming the error vector has zero expectation. It states that the following three statements are equivalent:

The rows of T lie in the row space of X.

The value of $\mathrm{T}\hat{\theta}$ does not depend on the particular solution $\hat{\theta}$ of the system of normal equations.

The function $\tau = \mathrm{T}\theta$ is estimable.

This lemma gives us a complete characterization of which parametric functions can be estimated without bias in the linear model.

Remark

We make no assumptions about the rank of the matrix X. Therefore, Lemma 5 holds even if the matrix $X^T X$ is singular.

Proof Strategy:

- We demonstrate the proof for the case $k = 1$; the general case follows similarly.
- The proof proceeds by showing a cyclic implication:

$$\textbf{(a)} \Rightarrow \textbf{(b)} \Rightarrow \textbf{(c)} \Rightarrow \textbf{(a)}.$$

Step 1: Proving (a) ⇒ (b)

- Assume **(a)**: $T^T \in \mathcal{L}(X^T X)$. This implies there exists a vector p such that $T^T = X^T X p$.
- For any solution $\widehat{\theta}$ of the system normal equations $X^T X\widehat{\theta} = X^T Y$, we have:

$$T\widehat{\theta} = (X^T X p)^T \widehat{\theta} = p^T X^T X\widehat{\theta} = p^T X^T Y.$$

- Since $p^T X^T Y$ does not depend on $\widehat{\theta}$, the value $T\widehat{\theta}$ is unique, satisfying condition **(b)**.

Comments

Let us now prove the lemma that establishes equivalence between three conditions of estimability for a linear parametric function of the form τ equals $T\theta$. We begin with an important observation: we do not assume anything about the rank of the matrix X. Thus, the lemma remains valid even when the matrix $X^T X$ is singular, that is, when the design matrix has linearly dependent columns.

We present the proof for the case when the number of components in τ equals one, that is, k equals one. The general case is proved analogously. The proof proceeds along a cycle of implications: condition (a) implies (b), which implies (c), which in turn implies (a).

First, suppose that the transpose of T lies in the column space of the matrix $X^T X$. This means there exists a vector p such that T^T equals $X^T X$ times p. Now, let $\hat{\theta}$ be any solution of the system of normal equations.

Then, T times $\hat{\theta}$ equals p^T times $X^T X$ times $\hat{\theta}$. But this simplifies to p^T times X^T times Y. Thus, the value of T times $\hat{\theta}$ does not depend on the particular choice of $\hat{\theta}$, which proves that condition (b) holds.

Proof of Lemma: Estimability Conditions (continued)

Step 2: Proving (b) ⇒ (c)

- For any λ such that $X^TX\lambda = 0$ we have that if $\widehat{\theta}$ is a solution of the system normal equations ($X^TX\widehat{\theta} = X^TY$) then $\bar{\theta} = \widehat{\theta} - \lambda$ is also the solution.
- Condition (b) implies $T\widehat{\theta} = T\bar{\theta}$, so we have that $0 = T\widehat{\theta} - T\bar{\theta} = T\lambda$
- since this equality holds for any λ which satisfies $X^TX\lambda = 0$, we have that $T \in \mathcal{L}(X^TX)$
- this means that there exist a vector p such that $T^T = X^TXp$ and

$$ET\widehat{\theta} = p^TX^TEY = p^TX^TX\theta = T\theta,$$

- Thus, $\tau = T\theta$ is an estimable function, satisfying condition (c).

Step 3: Proving (c) ⇒ (a)

- Suppose the function $\tau = T\theta$ is estimable.
- By definition, there exists a matrix A such that $\mathbb{E}[AY] = T\theta$.
- Thus we have $AX\theta = T\theta$, which implies $AX = T$ (see the proof of Lemma 3).
- Consequently, $T^T \in \mathcal{L}(X^T)$, which is equivalent to condition **(a)**.

The Lemma is proved. □

Comments

We will now show that condition (b) implies condition (c), and then we will complete the logical cycle by proving that condition (c) implies condition (a).

Assume condition (b) holds. This means that the value of T times $\hat{\theta}$ is the same for all solutions of the normal equation. Recall that the normal equation is $X^TX\hat{\theta} = X^TY$. If this system has more than one solution, then any other solution has the form $\hat{\theta}$ minus λ, where λ is a vector that satisfies $X^TX\lambda = 0$.

Now, since both $\hat{\theta}$ and $\hat{\theta}$ minus λ are valid solutions, and the value of T times $\hat{\theta}$ must be the same for both, we subtract the two expressions and get T times λ equals zero. And this must hold for any vector λ such that $X^TX\lambda$ equals zero.

This condition tells us that there exists a vector p such that T^T equals X^TX times p. In other words, T^T can be expressed as a linear combination of the columns of X^TX.

Now consider the expectation of T times $\hat{\theta}$. Since $\hat{\theta}$ is a solution to the normal equation, we have: the expectation of T times $\hat{\theta}$ equals p^T times X^T times the expectation of Y. Because the expectation of Y is equal to X times θ, this becomes: p^T times X^T times X times θ, which simplifies to T times θ.

This shows that T times θ is an unbiased linear estimator, so the function is estimable. Hence, condition (c) is satisfied.

To complete the proof, we now assume that τ equals $T\theta$ is estimable. Then by definition, there exists a matrix A such that the expected value of A times Y equals $T\theta$. This implies that A times X equals T. Taking transposes, we obtain that T^T belongs to the column space of X^T, which implies condition (a).

Thus, the equivalence of all three conditions is proven.

Theorem 2 (BLUE for τ)

If the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is non-singular and the standard assumptions of the linear regression model are satisfied, then for any matrix T of compatible dimension:

- The function $\tau = \mathrm{T}\theta$ is estimable.
- Its best linear unbiased estimator (BLUE) is

$$\widehat{\tau} = \mathrm{T}\widehat{\theta} = \mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}.$$

- The covariance matrix of the estimator $\widehat{\tau}$ is

$$\mathrm{D}_{\widehat{\tau}} = \sigma^2\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{T}^{\mathrm{T}}.$$

Model assumptions:

- The model is $\mathrm{Y} = \mathrm{X}\theta + \varepsilon$, where ε is a random error vector;
- $\mathbb{E}[\varepsilon] = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2\mathrm{I}$;
- The matrix X has full column rank: $\mathrm{rank}(\mathrm{X}) = \mathrm{k}$.

Comments

From the previously proven lemma and the Gauss–Markov theorem, we now obtain an important result concerning the estimation of linear functions of the parameter vector θ when the design matrix X has full column rank.

The theorem states the following: if the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is invertible and all classical assumptions of the linear regression model are satisfied, then for any matrix T of compatible dimensions, the linear function τ equals $\mathrm{T}\theta$ is estimable. Moreover, its best linear unbiased estimator — also known as the BLUE — is given by T times $\hat{\theta}$, where $\hat{\theta}$ equals $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$ is the usual least squares estimator.

The covariance matrix of this estimator is given by σ^2 times T times $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$ times T^{T}, which shows that the uncertainty in estimating τ depends both on the design matrix X and the matrix T.

It is important to note that this result relies on the invertibility of the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$, which corresponds to the assumption that the columns of X are linearly independent. This condition is necessary for the uniqueness of the least squares estimator. Otherwise, not all model parameters can be unbiased estimated.

In the next part, we will generalize this result to the case when the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is singular, that is, when the model is not of full rank.