Lecture Notes: Linear Combinations and Optimality in Linear Models

Undergraduate Mathematics Educator

Based on Lecture Transcript

Important Announcements and Quiz Information

- **Upcoming Quiz:**
  - **Date:** Monday, May 12th (Week 7).
  - **Time:** During the regular 2-hour lecture slot.
  - **Allowed Materials:** You may bring 4 pages of notes. This can be either 4 single-sided sheets OR 2 double-sided sheets.
  - **Format:** The notes MUST be printed on physical paper. Electronic devices (iPads, laptops, etc.) are **not permitted** for accessing notes during the quiz.
  - Further details will be posted in an official course announcement.
- Please ensure you review the material covered up to the end of Week 6 in preparation for the quiz.

1 Recap: The Linear Model and Least Squares Estimation

Let's begin by recalling the framework we've been working with: the linear model. We express the relationship between a response vector $\mathbf{Y}$ (of length n) and a set of predictor variables encoded in the design matrix $\mathbf{X}$ (size $n \times (p+1)$) via a parameter vector β (of length $p+1$) and an error term ϵ (of length n).

Definition 1.1 (Linear Model)**.** The linear model is given by:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

We make the following standard assumptions about the error term ϵ:

1. **Zero Mean:** $\mathrm{E}[\epsilon] = \mathbf{0}$ (a vector of zeros). This implies $\mathrm{E}[Y_i] = (\mathbf{X}\beta)_i$.
2. **Constant Variance and Uncorrelated Errors:** $\mathrm{Cov}(\epsilon) = \mathrm{E}[\epsilon\epsilon^T] = \sigma^2\mathbf{I}_n$, where $\sigma^2 > 0$ is the error variance and $\mathbf{I}_n$ is the $n \times n$ identity matrix. This means $\mathrm{Var}(\epsilon_i) = \sigma^2$ for all i, and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

The design matrix $\mathbf{X}$ is considered fixed and known. The parameters β and σ^2 are considered fixed but unknown constants that we wish to estimate or make inferences about.

Remark 1.2 (Alternative Notation)**.** Sometimes, we compactly write the assumptions on ϵ as $\epsilon \sim (0, \sigma^2\mathbf{I}_n)$. At this stage, this notation *only* refers to the mean and covariance structure (the first two moments). We have not yet assumed a specific distribution (like normality) for the errors.

Our primary tool for estimating β has been the method of least squares, which minimizes the sum of squared residuals $SSE = ||\mathbf{Y} - \mathbf{X}\beta||^2$.

Proposition 1.3 (Least Squares Estimators)**.**

- *The least squares estimator (LSE) for β is:*

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

  (assuming $\mathbf{X}^T\mathbf{X}$ is invertible, which usually holds if $\mathbf{X}$ has full column rank).

- *An unbiased estimator for the error variance σ^2 is:*

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-(p+1)} = \frac{||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2}{n-p-1}$$

  The denominator $n-p-1$ represents the degrees of freedom for error.

We established some fundamental properties of the LSE $\hat{\beta}$ last time:

Proposition 1.4 (Properties of $\hat{\beta}$)**.** *Under the linear model assumptions:*

- ***Unbiasedness:*** *$\hat{\beta}$ is an unbiased estimator of β, meaning $\mathrm{E}[\hat{\beta}] = \beta$.*
- ***Covariance Matrix:*** *The covariance matrix of the random vector $\hat{\beta}$ is*

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Proof Sketch Reminder.

- Unbiasedness: $\mathrm{E}[\hat{\beta}] = \mathrm{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathrm{E}[\mathbf{Y}]$. Since $\mathrm{E}[\mathbf{Y}] = \mathrm{E}[\mathbf{X}\beta + \epsilon] = \mathbf{X}\beta + \mathrm{E}[\epsilon] = \mathbf{X}\beta$, we get $\mathrm{E}[\hat{\beta}] = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\beta = \beta$.
- Covariance: We compute $\mathrm{Cov}(\hat{\beta}) = \mathrm{Cov}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y})$. Apply the property $\mathrm{Cov}(\mathbf{AY}) = \mathbf{A}\,\mathrm{Cov}(\mathbf{Y})\mathbf{A}^T$ for a constant matrix $\mathbf{A}$. Note that $\mathrm{Cov}(\mathbf{Y}) = \mathrm{Cov}(\mathbf{X}\beta + \epsilon) = \mathrm{Cov}(\epsilon) = \sigma^2\mathbf{I}_n$, since $\mathbf{X}\beta$ is constant. This gives:

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}) &= [(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T](\sigma^2\mathbf{I}_n)[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X})^{-1}]^T \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \quad \text{(since } (\mathbf{X}^T\mathbf{X})^{-1} \text{ is symmetric)} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}$$

■

2 Linear Combinations of Parameters

Often, we aren't interested in estimating the entire vector β, but rather a specific linear combination of its components. For example, we might want to estimate a single coefficient β_j, or the difference between two coefficients $\beta_i - \beta_j$, or the average effect $\frac{1}{p+1}\sum_{j=0}^{p}\beta_j$.

Definition 2.1 (Linear Combination)**.** Let $\mathbf{a} = (a_0, a_1, \dots, a_p)^T$ be a fixed, known vector of constants in $\mathbb{R}^{p+1}$. A **linear combination** of the parameters in β is defined as the scalar quantity:

$$\theta = \mathbf{a}^T\beta = \sum_{j=0}^{p} a_j\beta_j$$

Since β consists of unknown parameters, θ is also an unknown parameter.

Example 2.2 (Estimating a Single Coefficient)**.** Suppose we want to estimate the j-th coefficient, β_j (where the index j runs from 0 to p). We can achieve this by choosing the vector $\mathbf{a} = \mathbf{e}_j$, where $\mathbf{e}_j$ is the standard basis vector with a 1 in the $(j+1)$-th position (corresponding to β_j) and 0s elsewhere. Then,

$$\theta = \mathbf{e}_j^T \beta = \beta_j$$

So, individual coefficients are special cases of linear combinations.

Example 2.3 (Estimating the Average Coefficient)**.** Suppose we want to estimate the average of all coefficients. We can choose $\mathbf{a} = \frac{1}{p+1}(1, 1, \ldots, 1)^T = \frac{1}{p+1}\mathbf{1}$. Then,

$$\theta = \left(\frac{1}{p+1}\mathbf{1}\right)^T \beta = \frac{1}{p+1}\sum_{j=0}^{p} \beta_j = \bar{\beta}$$

This gives the average value of the regression coefficients.

Question 2.4. Given that we want to estimate $\theta = \mathbf{a}^T\beta$, and we already have a good estimator $\hat{\beta}$ for β, what would be a natural way to estimate θ?

The most intuitive approach is simply to replace the unknown β in the definition of θ with its estimator $\hat{\beta}$.

Definition 2.5 (Estimator for a Linear Combination)**.** The **least squares estimator** for the linear combination $\theta = \mathbf{a}^T\beta$ is given by:

$$\hat{\theta} = \mathbf{a}^T\hat{\beta}$$

Substituting the formula for $\hat{\beta}$, we get:

$$\hat{\theta} = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Remark 2.6 (Structure of $\hat{\theta}$)**.** Notice that $\hat{\theta}$ is a scalar value (since $\mathbf{a}$ is $(p+1) \times 1$ and $\hat{\beta}$ is $(p+1) \times 1$). It's also a random variable because it depends on the random vector $\mathbf{Y}$. Furthermore, we can see that $\hat{\theta}$ is a linear combination of the elements of $\mathbf{Y}$. Let's define a vector $\mathbf{c}$:

$$\mathbf{c} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}$$

Let's check the dimensions: $\mathbf{X}$ is $n \times (p+1)$, $(\mathbf{X}^T\mathbf{X})^{-1}$ is $(p+1) \times (p+1)$, and $\mathbf{a}$ is $(p+1) \times 1$. So, $\mathbf{c}$ is an $n \times 1$ vector. This vector $\mathbf{c}$ is fixed and known, as it only depends on $\mathbf{X}$ and our choice of $\mathbf{a}$. Then, we can rewrite $\hat{\theta}$ as:

$$\hat{\theta} = (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a})^T\mathbf{Y} = \mathbf{c}^T\mathbf{Y} = \sum_{i=1}^{n} c_i Y_i$$

This explicitly shows that $\hat{\theta}$ is a linear estimator (linear function of $\mathbf{Y}$).

Now, let's examine the properties of this natural estimator $\hat{\theta}$.

Proposition 2.7 (Properties of $\hat{\theta}$)**.** *The estimator $\hat{\theta} = \mathbf{a}^T\hat{\beta}$ has the following properties:*

- ***Unbiasedness:*** *$\hat{\theta}$ is an unbiased estimator of θ.*

$$\mathrm{E}[\hat{\theta}] = \theta$$

- ***Variance:*** *The variance of $\hat{\theta}$ is*

$$\mathrm{Var}(\hat{\theta}) = \sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}$$

Proof.

- **Unbiasedness:** Using the linearity of expectation and the unbiasedness of $\hat{\beta}$:

$$\mathrm{E}[\hat{\theta}] = \mathrm{E}[\mathbf{a}^T\hat{\beta}] = \mathbf{a}^T\,\mathrm{E}[\hat{\beta}] = \mathbf{a}^T\beta = \theta$$

- **Variance:** We can view the scalar $\hat{\theta}$ as a 1×1 random matrix. Its variance is then its 1×1 covariance matrix. Using the property $\mathrm{Cov}(\mathbf{AZ}) = \mathbf{A}\,\mathrm{Cov}(\mathbf{Z})\mathbf{A}^T$ with $\mathbf{A} = \mathbf{a}^T$ and $\mathbf{Z} = \hat{\beta}$:

$$\begin{aligned}\mathrm{Var}(\hat{\theta}) = \mathrm{Cov}(\mathbf{a}^T\hat{\beta}) &= \mathbf{a}^T\,\mathrm{Cov}(\hat{\beta})(\mathbf{a}^T)^T \\ &= \mathbf{a}^T[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{a} \\ &= \sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}\end{aligned}$$

■

Remark 2.8 (Alternative Variance Calculation)**.** We can also calculate the variance using the form $\hat{\theta} = \mathbf{c}^T\mathbf{Y}$, where $\mathbf{c} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}$.

$$\begin{aligned}\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\mathbf{c}^T\mathbf{Y}) &= \mathbf{c}^T\,\mathrm{Cov}(\mathbf{Y})\mathbf{c} \\ &= \mathbf{c}^T(\sigma^2\mathbf{I}_n)\mathbf{c} \\ &= \sigma^2\mathbf{c}^T\mathbf{c} \\ &= \sigma^2[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}]^T[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}] \\ &= \sigma^2\mathbf{a}^T[(\mathbf{X}^T\mathbf{X})^{-1}]^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} \\ &= \sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} \quad \text{(since } (\mathbf{X}^T\mathbf{X})^{-1} \text{ is symmetric)} \\ &= \sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}\end{aligned}$$

Reassuringly, both approaches yield the same result for the variance.

So, $\hat{\theta} = \mathbf{a}^T\hat{\beta}$ is a linear estimator (in $\mathbf{Y}$) and it's unbiased for θ. A natural question arises: Is this the *best* possible estimator we can find within a reasonable class? To answer this, we first need to define what "best" means.

3 Comparing Estimators: The Mean Squared Error Criterion

When comparing different estimators for the same parameter θ, a widely used criterion is the Mean Squared Error (MSE).

Definition 3.1 (Mean Squared Error (MSE))**.** Let $\hat{\theta}$ be any estimator for a parameter θ. The **Mean Squared Error** of $\hat{\theta}$ is defined as:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)^2]$$

The MSE measures the average squared distance between the estimator and the true parameter value. An estimator with a smaller MSE is generally preferred.

Remark 3.2 (Dependence of MSE on θ)**.** It's crucial to understand that the MSE of an estimator $\hat{\theta}$ is typically a function of the true, unknown parameter θ. The expectation $\mathrm{E}[\cdot]$ is taken with respect to the distribution of the data (which depends on θ), and θ itself appears in the expression.

Example 3.3 (Constant Estimator)**.** Consider the naive estimator $\hat{\theta}_{\text{const}} = 0$ for all data. If the true parameter happens to be $\theta = 0$, then $\mathrm{MSE}(\hat{\theta}_{\text{const}}) = \mathrm{E}[(0 - 0)^2] = 0$. This estimator is perfect in this specific case. However, if the true parameter is $\theta \neq 0$, then $\mathrm{MSE}(\hat{\theta}_{\text{const}}) = \mathrm{E}[(0 - \theta)^2] = \mathrm{E}[(-\theta)^2] = \theta^2$. The error grows quadratically as the true θ moves away from 0. This illustrates that the performance (MSE) of an estimator depends on the unknown true parameter value.

A fundamental result relates the MSE to two other important quantities: the variance and the bias of the estimator.

Definition 3.4 (Bias)**.** The **Bias** of an estimator $\hat{\theta}$ for θ is defined as:

$$\text{Bias}(\hat{\theta}) = \text{E}[\hat{\theta}] - \theta$$

An estimator is **unbiased** if $\text{Bias}(\hat{\theta}) = 0$ for all possible values of θ.

Theorem 3.5 (MSE Decomposition: Bias-Variance Tradeoff)**.** *The Mean Squared Error of an estimator $\hat{\theta}$ can be decomposed as:*

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$$

Proof. We start with the definition of MSE and employ the "add and subtract" trick involving $\text{E}[\hat{\theta}]$:

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \text{E}[(\hat{\theta} - \theta)^2]\\ &= \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}] + \text{E}[\hat{\theta}] - \theta)^2]\\ &= \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2 + (\text{E}[\hat{\theta}] - \theta)^2 + 2(\hat{\theta} - \text{E}[\hat{\theta}])(\text{E}[\hat{\theta}] - \theta)]\end{aligned}$$

Using the linearity of expectation:

$$\text{MSE}(\hat{\theta}) = \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2] + \text{E}[(\text{E}[\hat{\theta}] - \theta)^2] + \text{E}[2(\hat{\theta} - \text{E}[\hat{\theta}])(\text{E}[\hat{\theta}] - \theta)]$$

Let's examine each term:

- The first term is, by definition, the variance of $\hat{\theta}$: $\text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2] = \text{Var}(\hat{\theta})$.
- The second term involves $\text{E}[\hat{\theta}] - \theta$, which is the bias. Since θ is a constant and $\text{E}[\hat{\theta}]$ is also a constant (for a given θ), the bias is a constant. The expectation of a constant squared is just the constant squared: $\text{E}[(\text{E}[\hat{\theta}] - \theta)^2] = (\text{E}[\hat{\theta}] - \theta)^2 = [\text{Bias}(\hat{\theta})]^2$.
- For the third term (the cross-term), notice that $(\text{E}[\hat{\theta}] - \theta)$ is a constant factor. We can pull it out of the expectation:

$$\begin{aligned}\text{E}[2(\hat{\theta} - \text{E}[\hat{\theta}])(\text{E}[\hat{\theta}] - \theta)] &= 2(\text{E}[\hat{\theta}] - \theta)\,\text{E}[\hat{\theta} - \text{E}[\hat{\theta}]]\\ &= 2(\text{Bias}(\hat{\theta}))(\text{E}[\hat{\theta}] - \text{E}[\text{E}[\hat{\theta}]])\\ &= 2(\text{Bias}(\hat{\theta}))(\text{E}[\hat{\theta}] - \text{E}[\hat{\theta}]) \quad (\text{since } \text{E}[\hat{\theta}] \text{ is a constant})\\ &= 2(\text{Bias}(\hat{\theta}))(0) = 0\end{aligned}$$

Putting it all together:

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2 + 0$$

as required. ■

This decomposition is fundamental. It tells us that the MSE has two components:

1. **Variance:** How much the estimator $\hat{\theta}$ fluctuates around its own average value $\text{E}[\hat{\theta}]$.
2. **Squared Bias:** How far the average value of the estimator $\text{E}[\hat{\theta}]$ is from the true target θ.

Remark 3.6 (The Bias-Variance Tradeoff)**.** Unbiasedness (Bias = 0) seems desirable. If we shoot a rifle at a target (the true θ), unbiasedness means that, on average, our shots ($\hat{\theta}$) are centered on the target ($\mathrm{E}[\hat{\theta}] = \theta$). However, the MSE measures the average squared distance from the target, which depends on both the centering (bias) and the spread of the shots (variance). It's possible to have an estimator with a small bias but very large variance, leading to a large MSE. Conversely, sometimes introducing a small amount of bias might allow for a significant reduction in variance, potentially resulting in a lower overall MSE. This is the essence of the **bias-variance tradeoff**. Think of a rifle that isn't perfectly zeroed (biased) but shoots very tight groups (low variance). Its average squared error might be smaller than a perfectly zeroed rifle that sprays shots widely (unbiased but high variance).

In many classical statistical settings, including our current focus, we often restrict our attention to the class of **unbiased estimators**.

Corollary 3.7 (MSE for Unbiased Estimators)**.** *If $\hat{\theta}$ is an unbiased estimator of θ, then* $\mathrm{Bias}(\hat{\theta}) = 0$*, and its MSE simplifies to:*

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$$

Therefore, when comparing two unbiased estimators, the one with the smaller variance also has the smaller MSE and is considered "better" according to this criterion.

Our estimator $\hat{\theta} = \mathbf{a}^T\hat{\beta}$ is linear (in $\mathbf{Y}$) and unbiased. The question now becomes: Is there any *other* linear, unbiased estimator for θ that has a smaller variance than $\hat{\theta}$?

4 The Gauss-Markov Theorem: Optimality of Least Squares

The answer to the question above is provided by a cornerstone result in linear models theory: the Gauss-Markov Theorem. It establishes that within the class of linear unbiased estimators, the least squares estimator is optimal in the sense that it has the minimum variance.

Theorem 4.1 (Gauss-Markov Theorem)**.** *Consider the linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ with $\mathrm{E}[\epsilon] = \mathbf{0}$ and $\mathrm{Cov}(\epsilon) = \sigma^2\mathbf{I}_n$. Let $\theta = \mathbf{a}^T\beta$ be any linear combination of the parameters. Let $\hat{\theta} = \mathbf{a}^T\hat{\beta}$ be the least squares estimator of θ.*

*Then, $\hat{\theta}$ is the **Best Linear Unbiased Estimator (BLUE)** of θ. This means:*

1. ***Linear:** $\hat{\theta}$ is a linear function of $\mathbf{Y}$. (We already wrote $\hat{\theta} = \mathbf{c}^T\mathbf{Y}$).*

2. ***Unbiased:** $\hat{\theta}$ is unbiased for θ, i.e., $\mathrm{E}[\hat{\theta}] = \theta$. (We already proved this).*

3. ***Best:** Among all linear unbiased estimators of θ, $\hat{\theta}$ has the minimum variance. That is, if $\tilde{\theta}$ is any other estimator of θ such that $\tilde{\theta} = \tilde{\mathbf{c}}^T\mathbf{Y}$ for some constant vector $\tilde{\mathbf{c}}$ (linear) and $\mathrm{E}[\tilde{\theta}] = \theta$ (unbiased), then*

$$\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\tilde{\theta})$$

This inequality holds for all possible values of the true parameters β and $\sigma^2 > 0$.

Proof. Let $\hat{\theta} = \mathbf{c}^T\mathbf{Y}$ be the LSE, where $\mathbf{c} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}$. Let $\tilde{\theta}$ be any other linear unbiased estimator of θ. Since it's linear, it must be of the form $\tilde{\theta} = \tilde{\mathbf{c}}^T\mathbf{Y}$ for some fixed $n \times 1$ vector $\tilde{\mathbf{c}}$.

Define the difference vector $\delta = \tilde{\mathbf{c}} - \mathbf{c}$. Then $\tilde{\mathbf{c}} = \mathbf{c} + \delta$, and we can write:

$$\tilde{\theta} = (\mathbf{c} + \delta)^T\mathbf{Y} = \mathbf{c}^T\mathbf{Y} + \delta^T\mathbf{Y} = \hat{\theta} + \delta^T\mathbf{Y}$$

Now, let's use the unbiasedness condition for $\tilde{\theta}$: $\text{E}[\tilde{\theta}] = \theta$.

$$\begin{aligned}\theta = \text{E}[\tilde{\theta}] &= \text{E}[\hat{\theta} + \delta^T \mathbf{Y}] \\ &= \text{E}[\hat{\theta}] + \text{E}[\delta^T \mathbf{Y}] \\ &= \theta + \delta^T \, \text{E}[\mathbf{Y}] \quad (\text{since } \hat{\theta} \text{ is unbiased and } \delta \text{ is constant}) \\ &= \theta + \delta^T(\mathbf{X}\beta) \quad (\text{since } \; \text{E}[\mathbf{Y}] = \mathbf{X}\beta)\end{aligned}$$

For this equality $\theta = \theta + \delta^T \mathbf{X}\beta$ to hold for *all* possible parameter vectors β, we must have:

$$\delta^T \mathbf{X}\beta = 0 \quad \text{for all } \beta \in \mathbb{R}^{p+1} \tag{1}$$

The only way for this linear combination of the elements of β (with coefficients given by the row vector $\delta^T \mathbf{X}$) to be zero for all β is if the coefficient vector itself is the zero vector. That is:

$$\delta^T \mathbf{X} = \mathbf{0}^T \quad (\text{a } 1 \times (p+1) \text{ row vector of zeros}) \tag{2}$$

This condition $\delta^T \mathbf{X} = \mathbf{0}^T$ is a necessary consequence of requiring $\tilde{\theta}$ to be unbiased. It essentially means δ must be orthogonal to the column space of $\mathbf{X}$.

Now, let's compute the variance of $\tilde{\theta}$:

$$\begin{aligned}\text{Var}(\tilde{\theta}) &= \text{Var}(\hat{\theta} + \delta^T \mathbf{Y}) \\ &= \text{Var}(\hat{\theta}) + \text{Var}(\delta^T \mathbf{Y}) + 2\,\text{Cov}(\hat{\theta}, \delta^T \mathbf{Y}) \\ &= \text{Var}(\mathbf{c}^T \mathbf{Y}) + \text{Var}(\delta^T \mathbf{Y}) + 2\,\text{Cov}(\mathbf{c}^T \mathbf{Y}, \delta^T \mathbf{Y})\end{aligned}$$

We know $\text{Var}(\mathbf{c}^T \mathbf{Y}) = \text{Var}(\hat{\theta}) = \sigma^2 \mathbf{c}^T \mathbf{c}$. Similarly, $\text{Var}(\delta^T \mathbf{Y}) = \delta^T \,\text{Cov}(\mathbf{Y})\delta = \delta^T(\sigma^2 \mathbf{I}_n)\delta = \sigma^2 \delta^T \delta$. Now consider the covariance term:

$$\begin{aligned}\text{Cov}(\mathbf{c}^T \mathbf{Y}, \delta^T \mathbf{Y}) &= \mathbf{c}^T \,\text{Cov}(\mathbf{Y})\delta \\ &= \mathbf{c}^T(\sigma^2 \mathbf{I}_n)\delta \\ &= \sigma^2 \mathbf{c}^T \delta \\ &= \sigma^2 [\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{a}]^T \delta \\ &= \sigma^2 \mathbf{a}^T (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \delta)\end{aligned}$$

But from condition (2), we know $\delta^T \mathbf{X} = \mathbf{0}^T$, which implies $(\mathbf{X}^T \delta) = \mathbf{0}$. Therefore, the covariance term is zero:

$$\text{Cov}(\hat{\theta}, \delta^T \mathbf{Y}) = \sigma^2 \mathbf{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{0} = 0$$

Substituting these back into the variance expression for $\tilde{\theta}$:

$$\begin{aligned}\text{Var}(\tilde{\theta}) &= \text{Var}(\hat{\theta}) + \sigma^2 \delta^T \delta + 2(0) \\ &= \text{Var}(\hat{\theta}) + \sigma^2 ||\delta||^2\end{aligned}$$

Since $\sigma^2 > 0$ and $||\delta||^2 = ||\tilde{\mathbf{c}} - \mathbf{c}||^2 \geq 0$, we have $\sigma^2 ||\delta||^2 \geq 0$. Thus,

$$\text{Var}(\tilde{\theta}) \geq \text{Var}(\hat{\theta})$$

Equality holds if and only if $||\delta||^2 = 0$, which means $\delta = \mathbf{0}$, implying $\tilde{\mathbf{c}} = \mathbf{c}$, i.e., $\tilde{\theta} = \hat{\theta}$. This shows that the LSE $\hat{\theta}$ has the minimum variance among all linear unbiased estimators. ■

Remark 4.2. The crucial step where unbiasedness was used was in deriving $\delta^T \mathbf{X} = \mathbf{0}^T$. This condition was necessary to show that the covariance term between $\hat{\theta}$ and the difference term $\delta^T \mathbf{Y}$ is zero.

5 Summary and Looking Ahead

Let's consolidate what we've established:

- Under the standard linear model assumptions ($Y = X\beta + \epsilon, E[\epsilon] = 0, Cov(\epsilon) = \sigma^2 I$), the least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$ is unbiased for β, with $Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$.
- For any linear combination $\theta = a^T \beta$, the natural estimator $\hat{\theta} = a^T \hat{\beta}$ is also unbiased for θ, with $Var(\hat{\theta}) = \sigma^2 a^T (X^T X)^{-1} a$.
- We use Mean Squared Error (MSE) to compare estimators, where $MSE = Variance + Bias^2$.
- For unbiased estimators, minimizing MSE is equivalent to minimizing variance.
- The **Gauss-Markov Theorem** states that $\hat{\theta} = a^T \hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE) for θ, meaning it has the smallest variance among all estimators that are both linear functions of Y and unbiased for θ.

This provides strong justification for using the least squares estimator when our goal is point estimation and we value linearity and unbiasedness. However, our analysis so far has relied only on the first two moments (mean and covariance) of the error distribution. To proceed further into statistical inference – constructing confidence intervals, performing hypothesis tests – we need to make more specific assumptions about the *shape* of the error distribution.

The most common and mathematically tractable assumption is that the errors follow a normal distribution. This will allow us to determine the exact distributions of our estimators (not just their mean and variance) and build formal inferential procedures.

Before diving into that, let's briefly refresh our understanding of multivariate distributions, as our error term ϵ and our response vector Y are vectors.

6 Appendix: A Brief Reminder on Multivariate Distributions

When dealing with random vectors, like $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_m)^T$, we need concepts analogous to the CDF and PDF from the univariate case. Let's illustrate with the simplest multivariate case, $m = 2$.

Let $\mathbf{Z} = (Z_1, Z_2)^T$ be a 2-dimensional random vector. Z_1 and Z_2 are random variables defined on the same probability space, so they have a joint behavior.

Definition 6.1 (Joint Cumulative Distribution Function (CDF))**.** The **joint CDF** of $\mathbf{Z} = (Z_1, Z_2)^T$ is a function $F_{\mathbf{Z}} : \mathbb{R}^2 \to [0, 1]$ defined as:

$$F_{\mathbf{Z}}(z_1, z_2) = P(Z_1 \le z_1, Z_2 \le z_2)$$

for any point $(z_1, z_2) \in \mathbb{R}^2$. Geometrically, $F_{\mathbf{Z}}(z_1, z_2)$ gives the probability mass accumulated in the quadrant to the "south-west" of the point (z_1, z_2).

Just as in the univariate case, if the CDF is sufficiently smooth, we can define a density function.

Definition 6.2 (Joint Probability Density Function (PDF))**.** If the joint CDF $F_{\mathbf{Z}}(z_1, z_2)$ is differentiable, the **joint PDF** of $\mathbf{Z} = (Z_1, Z_2)^T$ is a function $f_{\mathbf{Z}} : \mathbb{R}^2 \to [0, \infty)$ defined by the mixed partial derivative:

$$f_{\mathbf{Z}}(z_1, z_2) = \frac{\partial^2 F_{\mathbf{Z}}(z_1, z_2)}{\partial z_1 \partial z_2}$$

(provided the derivative exists). The PDF must satisfy $f_{\mathbf{Z}}(z_1, z_2) \geq 0$ for all (z_1, z_2), and its integral over the whole plane must be 1:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\mathbf{Z}}(z_1, z_2) dz_1 dz_2 = 1$$

Probabilities are found by integrating the PDF over regions: $P((Z_1, Z_2) \in A) = \iint_A f_{\mathbf{Z}}(z_1, z_2) dz_1 dz_2$.

Remark 6.3. When the PDF exists, the CDF and PDF provide equivalent information about the distribution of the random vector. Just as in the univariate case, knowing one allows you to determine the other. These concepts generalize naturally to m dimensions, involving m-variate integrals and m-th order mixed partial derivatives.

Next time, we will introduce the *multivariate normal distribution*, which will be our key distributional assumption for ϵ moving forward.