Lecture Notes: Multidimensional Random Vectors and Introduction to the Linear Model

Undergraduate Math Educator

Based on lecture given post-break, covering weeks prior

Administrative Announcements

A few key dates and details to keep in mind:

- **Midterm Quiz:** Please note that we have a midterm quiz scheduled for **May 12th**. That's coming up in about three weeks.
- **Quiz Material:** The material covered will likely be up to the topics discussed one week prior, meaning content covered up to **May 5th**.
- **Review Session:** I will try my best, though I can't promise, to hold a dedicated review session during the tutorial the week before the quiz. This would be separate from the regular tutorial content.
- **Moed Bet (Second Chance):** There will be a second chance date (*Moed Bet*) for the quiz, but please be aware this is **only for students who are eligible** according to the rules (typically for excused absences, etc.). Do not plan on this being a general retake opportunity. If you are unable to attend the main quiz for a valid reason, you may be eligible for a special subsequent date.
- **Office Hours:** Please make use of office hours if you have questions about the material or the upcoming quiz. Statistics can be tricky, but we're here to help!
- **New Homework Assignment:** Have you started the new homework assignment (the "additional exercise sheet")? How did you find it, especially question 4 onwards? It's good practice for the concepts we're discussing.

Let's make sure we're all on track. Okay, let's begin with the mathematics.

1 Recap and Deep Dive: Multidimensional Random Vectors

Over the last week before the break, and continuing now, we've been exploring the fascinating world of **multidimensional random variables**, often represented as random vectors. This is a crucial extension of the single-variable probability concepts you're familiar with. We also started touching upon the assumptions underpinning the **linear model**, which will be a central theme in this course.

Today, we'll solidify our understanding of random vectors and their properties in the first part, and then delve deeper into the assumptions of the linear model in the second part.

1.1 Random Vectors and Expectation

Let's start with the basics. Imagine we have n random variables, $Z_1, Z_2, \ldots, Z_n$. We can group these together into a column vector:

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} \in \mathbb{R}^n$$

This Z is what we call a **random vector**. Each component Z_i is itself a random variable.

Definition 1.1 (Expectation of a Random Vector)**.** *The **expectation** (or expected value) of a random vector $Z \in \mathbb{R}^n$ is defined as the vector of the expectations of its individual components:*

$$\mathbb{E}[Z] = \begin{pmatrix} \mathbb{E}[Z_1] \\ \mathbb{E}[Z_2] \\ \vdots \\ \mathbb{E}[Z_n] \end{pmatrix} \in \mathbb{R}^n$$

Each $\mathbb{E}[Z_i]$ here is the standard, or marginal, expectation of the random variable Z_i, considered on its own (not conditional on the other components).

Remark 1.2. Just like with scalar random variables, the expectation represents the "center of mass" or the long-run average of the random vector.

Many properties of expectation carry over naturally from the scalar case.

Proposition 1.3 (Linearity of Expectation for Vectors/Matrices)**.** *Let Z and W be random vectors of the same dimension, and let A, B, and C be constant (non-random) matrices or vectors of compatible dimensions. Then:*

1. $\mathbb{E}[Z + W] = \mathbb{E}[Z] + \mathbb{E}[W]$
2. $\mathbb{E}[AZ] = A\mathbb{E}[Z]$
3. $\mathbb{E}[AZ + C] = A\mathbb{E}[Z] + C$
4. $\mathbb{E}[AZB] = A\mathbb{E}[Z]B$ *(if dimensions allow multiplication)*

We need to be careful with matrix multiplication order, but the linearity principle holds. Proving these relies on applying the definition component-wise.

1.2 The Variance-Covariance Matrix

While expectation tells us about the center, the **variance-covariance matrix** (often just called the covariance matrix) tells us about the spread and the linear relationships between the components of a random vector.

Definition 1.4 (Covariance Matrix of Two Random Vectors)**.** *Let $Z \in \mathbb{R}^n$ and $W \in \mathbb{R}^p$ be random vectors. Their **covariance matrix** is defined as:*

$$\text{Cov}(Z, W) = \mathbb{E}\left[(Z - \mathbb{E}[Z])(W - \mathbb{E}[W])^T\right]$$

This results in an $n \times p$ matrix. The (i, j)-th entry of this matrix is $\text{Cov}(Z_i, W_j) = \mathbb{E}[(Z_i - \mathbb{E}[Z_i])(W_j - \mathbb{E}[W_j])]$.

Remark 1.5. Think about the dimensions: $(Z - \mathbb{E}[Z])$ is $n \times 1$, and $(W - \mathbb{E}[W])^T$ is $1 \times p$. Their outer product gives an $n \times p$ matrix. The expectation is taken element-wise.

Definition 1.6 (Variance-Covariance Matrix of a Single Random Vector)**.** *A particularly important case is the covariance of a random vector $Z \in \mathbb{R}^n$ with itself. This is called the **variance-covariance matrix** of Z, often denoted* $\mathrm{Var}(Z)$*:*

$$\mathrm{Var}(Z) = \mathrm{Cov}(Z, Z) = \mathbb{E}\left[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^T\right]$$

This is always an $n \times n$ square matrix.

What does this matrix look like?

$$\mathrm{Var}(Z) = \begin{pmatrix} \mathrm{Cov}(Z_1, Z_1) & \mathrm{Cov}(Z_1, Z_2) & \cdots & \mathrm{Cov}(Z_1, Z_n) \\ \mathrm{Cov}(Z_2, Z_1) & \mathrm{Cov}(Z_2, Z_2) & \cdots & \mathrm{Cov}(Z_2, Z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(Z_n, Z_1) & \mathrm{Cov}(Z_n, Z_2) & \cdots & \mathrm{Cov}(Z_n, Z_n) \end{pmatrix}$$

Notice that the diagonal entries are $\mathrm{Cov}(Z_i, Z_i) = \mathrm{Var}(Z_i)$, the variances of the individual components. The off-diagonal entries $\mathrm{Cov}(Z_i, Z_j)$ (for $i \neq j$) are the covariances between pairs of components, measuring their linear association.

1.3 Properties of Covariance Matrices

Covariance matrices have several crucial properties, extending familiar scalar properties:

Proposition 1.7 (Properties of Covariance)**.** *Let $Z \in \mathbb{R}^n$, $W \in \mathbb{R}^p$, $R \in \mathbb{R}^n$ be random vectors, and let A, B, C be constant matrices/vectors of compatible dimensions.*

1. ***Symmetry of Variance:*** $\mathrm{Var}(Z)$ *is a symmetric matrix. (*$\mathrm{Var}(Z) = \mathrm{Var}(Z)^T$*)*
2. ***Relationship between Cov(Z, W) and Cov(W, Z):*** $\mathrm{Cov}(Z, W) = \mathrm{Cov}(W, Z)^T$. *(This relates the $n \times p$ matrix to the $p \times n$ matrix).*
3. ***Bilinearity:***
   - $\mathrm{Cov}(Z + R, W) = \mathrm{Cov}(Z, W) + \mathrm{Cov}(R, W)$
   - $\mathrm{Cov}(Z, W + S) = \mathrm{Cov}(Z, W) + \mathrm{Cov}(Z, S)$ *(where $S \in \mathbb{R}^p$)*
4. ***Effect of Linear Transformations:***
   - $\mathrm{Cov}(AZ, BW) = A\mathrm{Cov}(Z, W)B^T$
   - $\mathrm{Var}(AZ) = \mathrm{Cov}(AZ, AZ) = A\mathrm{Cov}(Z, Z)A^T = A\mathrm{Var}(Z)A^T$
   - $\mathrm{Var}(Z + C) = \mathrm{Var}(Z)$ *(Adding a constant doesn't change variance/covariance)*
5. ***Positive Semi-Definiteness:*** *For any random vector Z, its variance-covariance matrix* $\mathrm{Var}(Z)$ *is positive semi-definite (PSD). This means that for any constant vector $v \in \mathbb{R}^n$, the scalar quantity* $v^T\mathrm{Var}(Z)v \geq 0$.

Remark 1.8 (Intuition for PSD)**.** Why must $\mathrm{Var}(Z)$ be PSD? Consider the scalar random variable $Y = v^T Z = \sum v_i Z_i$, which is a linear combination of the components of Z. Its variance is $\mathrm{Var}(Y)$. Using the transformation property (4b from Prop 1.7) with $A = v^T$ (a $1 \times n$ matrix), we get $\mathrm{Var}(v^T Z) = v^T\mathrm{Var}(Z)(v^T)^T = v^T\mathrm{Var}(Z)v$. Since the variance of any scalar random variable must be non-negative, it follows that $v^T\mathrm{Var}(Z)v \geq 0$ for all v. This property is fundamental!

Remark 1.9 (Independence and Covariance)**.** If components Z_i and Z_j are independent, then $\mathrm{Cov}(Z_i, Z_j) = 0$. If all components of Z are mutually independent, then $\mathrm{Var}(Z)$ will be a diagonal matrix. However, the converse is not generally true: $\mathrm{Cov}(Z_i, Z_j) = 0$ (uncorrelated) does not imply independence, *except* in the special case of normally distributed random variables.

Let's work through an example to make this concrete.

Example 1.10 (Standard Normal Vector)**.** Let $Z_1, Z_2, \ldots, Z_5$ be independent random variables, each following a standard normal distribution, $Z_i \sim N(0, 1)$. Let $Z = (Z_1, \ldots, Z_5)^T$. Find $\mathbb{E}[Z]$ and $\mathrm{Var}(Z)$.

Solution:

- **Expectation:** Since $Z_i \sim N(0, 1)$, we know $\mathbb{E}[Z_i] = 0$ for all $i = 1, \ldots, 5$. Therefore,

$$\mathbb{E}[Z] = \begin{pmatrix} \mathbb{E}[Z_1] \\ \vdots \\ \mathbb{E}[Z_5] \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}_5 \in \mathbb{R}^5$$

  The expectation is the zero vector in $\mathbb{R}^5$.

- **Variance-Covariance Matrix:** We need to find the 5×5 matrix $\mathrm{Var}(Z)$. The entry (i, j) is $\mathrm{Cov}(Z_i, Z_j)$.
  - For the diagonal entries ($i = j$): $\mathrm{Cov}(Z_i, Z_i) = \mathrm{Var}(Z_i)$. Since $Z_i \sim N(0, 1)$, we have $\mathrm{Var}(Z_i) = 1$.
  - For the off-diagonal entries ($i \neq j$): $\mathrm{Cov}(Z_i, Z_j)$. Since Z_i and Z_j are independent for $i \neq j$, their covariance is 0. $\mathrm{Cov}(Z_i, Z_j) = 0$.

  Putting this together, the matrix has 1s on the diagonal and 0s everywhere else. This is the identity matrix of size 5.

$$\mathrm{Var}(Z) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = I_5$$

So, for a vector of independent standard normal variables, the expectation is the zero vector and the covariance matrix is the identity matrix. This is a foundational result.

Example 1.11 (Affine Transformation of a Random Vector)**.** Continuing with $Z \in \mathbb{R}^5$ from Example 1.10, where $\mathbb{E}[Z] = \mathbf{0}_5$ and $\mathrm{Var}(Z) = I_5$. Define a transformation $H : \mathbb{R}^5 \to \mathbb{R}^3$ by $Y = H(Z) = BZ + c$, where

$$B = \begin{pmatrix} 2 & -3 & 4 & 0 & 0 \\ 0 & 8 & 0 & -4 & -6 \\ 0 & 0 & 0 & 6 & 1 \end{pmatrix} \quad \text{and} \quad c = \begin{pmatrix} 7 \\ 3 \\ -1 \end{pmatrix}$$

1. Is the transformation H linear?
2. Calculate the expectation vector $\mathbb{E}[Y]$ and the covariance matrix $\mathrm{Var}(Y)$.

Solution:

1. **Linearity of H:** A transformation H is linear if $H(\mathbf{0}) = \mathbf{0}$ and $H(\alpha v_1 + \beta v_2) = \alpha H(v_1) + \beta H(v_2)$. Let's check the first condition:

$$H(\mathbf{0}_5) = B \cdot \mathbf{0}_5 + c = \mathbf{0}_3 + c = c = \begin{pmatrix} 7 \\ 3 \\ -1 \end{pmatrix}$$

Since $H(\mathbf{0}) \neq \mathbf{0}_3$, the transformation H is **not linear**. It is an *affine* transformation, which is a linear transformation ($Z \mapsto BZ$) followed by a translation (adding c).

2. **Expectation and Variance of Y:** We use the properties of expectation and variance under linear transformations (Prop 1.3 and 1.7).

- **Expectation:**

$$\mathbb{E}[Y] = \mathbb{E}[BZ + c] = \mathbb{E}[BZ] + \mathbb{E}[c]$$

Using linearity and the fact that c is constant:

$$\mathbb{E}[Y] = B\mathbb{E}[Z] + c$$

We know $\mathbb{E}[Z] = \mathbf{0}_5$, so:

$$\mathbb{E}[Y] = B \cdot \mathbf{0}_5 + c = \mathbf{0}_3 + c = c = \begin{pmatrix} 7 \\ 3 \\ -1 \end{pmatrix}$$

The expected value of Y is simply the translation vector c.

- **Variance-Covariance Matrix:**

$$\text{Var}(Y) = \text{Var}(BZ + c)$$

Adding a constant vector does not affect the variance:

$$\text{Var}(Y) = \text{Var}(BZ)$$

Using the property $\text{Var}(AZ) = A\text{Var}(Z)A^T$:

$$\text{Var}(Y) = B\text{Var}(Z)B^T$$

We know $\text{Var}(Z) = I_5$, so:

$$\text{Var}(Y) = BI_5B^T = BB^T$$

Now we need to compute the matrix product BB^T:

$$B^T = \begin{pmatrix} 2 & 0 & 0 \\ -3 & 8 & 0 \\ 4 & 0 & 0 \\ 0 & -4 & 6 \\ 0 & -6 & 1 \end{pmatrix}$$

$$BB^T = \begin{pmatrix} 2 & -3 & 4 & 0 & 0 \\ 0 & 8 & 0 & -4 & -6 \\ 0 & 0 & 0 & 6 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ -3 & 8 & 0 \\ 4 & 0 & 0 \\ 0 & -4 & 6 \\ 0 & -6 & 1 \end{pmatrix}$$

Performing the multiplication (row by column):

$$
\begin{aligned}
(1,1) :& (2)(2) + (-3)(-3) + (4)(4) + 0 + 0 = 4 + 9 + 16 = 29 \\
(1,2) :& (2)(0) + (-3)(8) + (4)(0) + 0 + 0 = -24 \\
(1,3) :& (2)(0) + (-3)(0) + (4)(0) + (0)(6) + (0)(1) = 0 \\
(2,1) :& (0)(2) + (8)(-3) + (0)(4) + (-4)(0) + (-6)(0) = -24 \\
(2,2) :& (0)(0) + (8)(8) + (0)(0) + (-4)(-4) + (-6)(-6) = 64 + 16 + 36 = 116 \\
(2,3) :& (0)(0) + (8)(0) + (0)(0) + (-4)(6) + (-6)(1) = -24 - 6 = -30 \\
(3,1) :& 0 + 0 + 0 + 0 + 0 = 0 \\
(3,2) :& 0 + 0 + 0 + (6)(-4) + (1)(-6) = -24 - 6 = -30 \\
(3,3) :& 0 + 0 + 0 + (6)(6) + (1)(1) = 36 + 1 = 37
\end{aligned}
$$

So, the covariance matrix is:

$$\mathrm{Var}(Y) = BB^T = \begin{pmatrix} 29 & -24 & 0 \\ -24 & 116 & -30 \\ 0 & -30 & 37 \end{pmatrix}$$

Note: As expected, $\mathrm{Var}(Y)$ is a 3×3 symmetric matrix.

This example illustrates how to handle affine transformations and highlights that the non-linear shift c only affects the mean, not the covariance structure.

Remark 1.12 (Independence from Distribution Type for Mean/Variance)**.** Would the results for $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$ in Example 1.11 change if the Z_i were independent $U(-\sqrt{3}, \sqrt{3})$ variables instead of $N(0,1)$? Let's check the properties of $Z_i \sim U(-\sqrt{3}, \sqrt{3})$:

- $\mathbb{E}[Z_i] = \frac{-\sqrt{3}+\sqrt{3}}{2} = 0$.
- $\mathrm{Var}(Z_i) = \frac{(\sqrt{3}-(-\sqrt{3}))^2}{12} = \frac{(2\sqrt{3})^2}{12} = \frac{12}{12} = 1$.

Since the Z_i are still independent, $\mathbb{E}[Z_i] = 0$, and $\mathrm{Var}(Z_i) = 1$, the input vector Z still has $\mathbb{E}[Z] = \mathbf{0}_5$ and $\mathrm{Var}(Z) = I_5$. The calculations for $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$ relied *only* on the mean vector and covariance matrix of Z, not on the specific type of distribution (Normal vs. Uniform). Therefore, the results would be exactly the same. This is an important point: many properties and calculations in linear models depend only on first and second moments (mean, variance, covariance), not the full distribution. We often denote this assumption generically as $Z \sim (\mathbf{0}, I)$, meaning a random vector with mean $\mathbf{0}$ and covariance I.

2 Comparing Covariance Matrices

Sometimes we want to compare the ”spread” or ”variability” represented by two different covariance matrices. This leads to the concept of ordering matrices in terms of positive semi-definiteness.

Theorem 2.1 (Equivalence Conditions for Covariance Dominance)**.** *Let $Z, W \in \mathbb{R}^p$ be random vectors. The following statements are equivalent:*

1. *For every constant (non-random) vector $v \in \mathbb{R}^p$, $\mathrm{Var}(v^T Z) \geq \mathrm{Var}(v^T W)$.*
2. *The matrix $B = \mathrm{Var}(Z) - \mathrm{Var}(W)$ is positive semi-definite ($B \succeq 0$).*

3. *There exists a matrix C such that $B = \text{Var}(Z) - \text{Var}(W) = CC^T$. (This relates to the existence of a "matrix square root" or Cholesky factor for PSD matrices).*

Sketch of Equivalence 1 $\iff$ 2. We established the core relationship in Remark 1.8: $\text{Var}(v^T Z) = v^T \text{Var}(Z) v$ and $\text{Var}(v^T W) = v^T \text{Var}(W) v$. Let $B = \text{Var}(Z) - \text{Var}(W)$. Then, $v^T B v = v^T(\text{Var}(Z) - \text{Var}(W))v = v^T \text{Var}(Z) v - v^T \text{Var}(W) v = \text{Var}(v^T Z) - \text{Var}(v^T W)$.

Now the equivalence is clear:

- Assume (1) holds: $\text{Var}(v^T Z) \geq \text{Var}(v^T W)$ for all v. This implies $\text{Var}(v^T Z) - \text{Var}(v^T W) \geq 0$ for all v. Therefore, $v^T B v \geq 0$ for all v, which means B is PSD by definition. So, (1) $\Longrightarrow$ (2).
- Assume (2) holds: B is PSD, meaning $v^T B v \geq 0$ for all v. This implies $Var(v^T Z) - \text{Var}(v^T W) \geq 0$ for all v. Therefore, $\text{Var}(v^T Z) \geq \text{Var}(v^T W)$ for all v. So, (2) $\Longrightarrow$ (1).

The equivalence between (2) and (3) relates to properties of positive semi-definite matrices. Since $B = \text{Var}(Z) - \text{Var}(W)$ must be symmetric (as $\text{Var}(Z)$ and $\text{Var}(W)$ are symmetric), if B is PSD, such a C exists (e.g., from Cholesky decomposition or eigenvalue decomposition). Conversely, if $B = CC^T$, then $v^T B v = v^T C C^T v = (C^T v)^T (C^T v) = \|C^T v\|^2 \geq 0$, so B is PSD. □

Remark 2.2 (Motivation for Theorem 2.1)**.** Why is this theorem important? In statistics, we often compare estimators. Suppose $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are two different vector estimators for some true parameter vector $\boldsymbol{\beta}$. Statement (1) says that for any linear combination of the parameters $v^T \boldsymbol{\beta}$, the variance of the estimate using $\hat{\boldsymbol{\beta}}_1$ ($v^T \hat{\boldsymbol{\beta}}_1$) is no smaller than the variance using $\hat{\boldsymbol{\beta}}_2$ ($v^T \hat{\boldsymbol{\beta}}_2$). Statement (2) provides a potentially easier way to check this condition by just examining the difference of the covariance matrices. If $\text{Var}(\hat{\boldsymbol{\beta}}_1) - \text{Var}(\hat{\boldsymbol{\beta}}_2)$ is PSD, we say that $\hat{\boldsymbol{\beta}}_2$ is "more efficient" than $\hat{\boldsymbol{\beta}}_1$ in this matrix sense. We will see later that the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ is the "Best Linear Unbiased Estimator" (BLUE), meaning it has the smallest variance (in this matrix sense) among all linear unbiased estimators. This theorem provides the mathematical language for such comparisons.

3 Introduction to the Linear Model

Now we transition to the core topic of this course: the linear model.

3.1 What is a Statistical Model?

Remember from your introductory courses, a statistical model is a set of assumptions about how data are generated. Often, we assume the data $Y_1, \ldots, Y_n$ come from a specific family of distributions characterized by some unknown parameter θ (e.g., $Y_i \sim \text{Exponential}(\lambda)$, where $\theta = \lambda$). The goal of statistical inference is typically to estimate θ from the observed data.

The linear model provides a framework for modeling the relationship between an outcome variable Y and one or more predictor variables (or covariates) $X_1, \ldots, X_p$. It assumes a specific, linear, structure for this relationship, plus some assumptions about the randomness involved.

3.2 The Linear Model Formulation

Let Y_i be the outcome variable for the i-th observation ($i = 1, \ldots, n$). Let $x_{i1}, \ldots, x_{ip}$ be the values of p predictor variables for the i-th observation. We often include an intercept term by setting $x_{i0} = 1$ for all i.

We can organize this data:

- Outcome vector: $Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$
- Design Matrix: X is an $n \times (p+1)$ matrix where the i-th row is $(x_{i0}, x_{i1}, \dots, x_{ip})$.

$$X = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

(Assuming an intercept is included).

The **linear model** postulates that the relationship between Y and X can be described as:

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

where:

- $Y \in \mathbb{R}^n$ is the random vector of outcomes.
- $X \in \mathbb{R}^{n \times (p+1)}$ is the **design matrix**, usually treated as **fixed and known**. (Or conditions are conditional on X).
- $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ is the vector of unknown **regression coefficients** (parameters). $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$. These are fixed, unknown constants we want to estimate.
- $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the vector of random **errors** or disturbances. $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$.

The randomness in Y comes entirely from the error term $\boldsymbol{\epsilon}$. The term $X\boldsymbol{\beta}$ represents the systematic part of the relationship. The error term ϵ_i captures all other factors influencing Y_i that are not included in the predictors X, including inherent randomness. If $\boldsymbol{\epsilon}$ were zero, Y would be perfectly determined by X.

3.3 Assumptions of the Classical Linear Model

For the standard theory (especially Ordinary Least Squares - OLS) to work nicely, we typically make the following assumptions about the error term $\boldsymbol{\epsilon}$, often called the **Gauss-Markov assumptions**:

1. **Linearity:** The relationship $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ holds. (This is the model definition itself).
2. **Zero Conditional Mean (Strict Exogeneity):** $\mathbb{E}[\boldsymbol{\epsilon}|X] = \mathbf{0}_n$. This implies $\mathbb{E}[\epsilon_i|X] = 0$ for all i. It also implies the unconditional mean is zero: $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbb{E}[\mathbb{E}[\boldsymbol{\epsilon}|X]] = \mathbb{E}[\mathbf{0}] = \mathbf{0}$. *Intuition:* The errors average out to zero regardless of the values of the predictors. There's no systematic tendency for the error to be positive or negative for particular X values. A violation might occur if the true relationship is non-linear but we fit a linear model. For example, if the true relationship is $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$, but we fit $Y = \gamma_0 + \gamma_1 X + \delta$, the new error term δ might have a mean that depends on X.
3. **Homoscedasticity and No Correlation:** The conditional covariance matrix of the errors, given X, is constant and diagonal:

$$\text{Var}(\boldsymbol{\epsilon}|X) = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|X] = \sigma^2 I_n$$

where σ^2 is a positive, unknown constant scalar (the error variance). This single matrix assumption combines two ideas:

- **Homoscedasticity (Constant Variance):** $\text{Var}(\epsilon_i|X) = \sigma^2$ for all i. The spread of the errors is the same for all observations, regardless of their X values.
- **No Correlation:** $\text{Cov}(\epsilon_i, \epsilon_j|X) = 0$ for all $i \neq j$. The errors for different observations are uncorrelated.

Intuition: Homoscedasticity means the variability of the outcome around the regression line is constant. Imagine plotting residuals against fitted values: homoscedasticity looks like a random horizontal band. Heteroscedasticity (non-constant variance) might appear as a funnel shape (variance increasing with fitted value) or other systematic patterns. No correlation means knowing the error for one observation gives no information about the error for another. This might be violated with time series data (errors today correlated with errors yesterday) or clustered data (e.g., student scores within the same school might be correlated).

4. **(Often Implicit) Fixed Design or Full Rank:** The design matrix X is treated as fixed (non-random) or conditions are conditional on X. We also usually assume X has full column rank $(p+1)$, meaning no perfect multicollinearity and $n \geq p+1$. This ensures (X^TX) is invertible.

5. **(For Inference) Normality:** Sometimes, for hypothesis testing and confidence intervals, we add the assumption that the errors are normally distributed: $\boldsymbol{\epsilon}|X \sim N(\mathbf{0}_n, \sigma^2 I_n)$. This assumption is *not* needed for OLS to be unbiased or for the variance formula to hold, but it simplifies distributional results for tests and intervals.

We can summarize assumptions (2) and (3) concisely using the notation we developed earlier:

$$\boldsymbol{\epsilon}|X \sim (\mathbf{0}_n, \sigma^2 I_n)$$

This notation implies the mean vector is $\mathbf{0}_n$ and the covariance matrix is $\sigma^2 I_n$, without necessarily assuming normality unless stated.

3.4 Consequences of Assumptions for Y

Under these assumptions, what can we say about the outcome vector Y?

- **Conditional Mean of Y:**

  $$\mathbb{E}[Y|X] = \mathbb{E}[X\boldsymbol{\beta} + \boldsymbol{\epsilon}|X] = \mathbb{E}[X\boldsymbol{\beta}|X] + \mathbb{E}[\boldsymbol{\epsilon}|X]$$

  Since X and $\boldsymbol{\beta}$ are fixed (or conditioned upon), $\mathbb{E}[X\boldsymbol{\beta}|X] = X\boldsymbol{\beta}$. Using Assumption 2, $\mathbb{E}[\boldsymbol{\epsilon}|X] = \mathbf{0}$.

  $$\mathbb{E}[Y|X] = X\boldsymbol{\beta} + \mathbf{0} = X\boldsymbol{\beta}$$

  The conditional expectation of the outcome lies exactly on the regression line/plane defined by $X\boldsymbol{\beta}$.

- **Conditional Variance of Y:**

  $$\text{Var}(Y|X) = \text{Var}(X\boldsymbol{\beta} + \boldsymbol{\epsilon}|X)$$

  Adding a constant ($X\boldsymbol{\beta}$ is constant given X) doesn't change variance:

  $$\text{Var}(Y|X) = \text{Var}(\boldsymbol{\epsilon}|X)$$

  Using Assumption 3:

  $$\text{Var}(Y|X) = \sigma^2 I_n$$

  The variability of Y around its conditional mean is determined solely by the error variance σ^2, and the outcomes Y_i are conditionally uncorrelated with the same conditional variance.

4 The Ordinary Least Squares (OLS) Estimator

Our goal is to estimate the unknown parameter vector $\boldsymbol{\beta}$. The most common method is Ordinary Least Squares (OLS).

Definition 4.1 (OLS Estimator)**.** *The OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ (often denoted simply $\hat{\boldsymbol{\beta}}$) is the vector that minimizes the sum of squared residuals (SSR):*

$$SSR(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - x_i^T \boldsymbol{\beta})^2 = \|Y - X\boldsymbol{\beta}\|^2$$

*where x_i^T is the i-th row of X. The vector $\hat{\boldsymbol{\beta}}$ that minimizes this is given by the solution to the **normal equations** $X^T(Y - X\hat{\boldsymbol{\beta}}) = \mathbf{0}$, which yields:*

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

(Assuming $X^T X$ is invertible, which requires X to have full column rank, meaning no perfect multicollinearity among predictors and $n \geq p + 1$).

Remark 4.2. Finding the OLS estimator is a mathematical optimization problem that does *not* require any of the statistical assumptions (Assumptions 1-3) about the error term. It's simply finding the vector $\hat{\boldsymbol{\beta}}$ such that the fitted values $X\hat{\boldsymbol{\beta}}$ are the orthogonal projection of Y onto the column space of X. However, the *properties* of this estimator (like unbiasedness and its variance) rely heavily on those assumptions.

4.1 Properties of the OLS Estimator

Let's investigate the properties of $\hat{\boldsymbol{\beta}}$ under the Gauss-Markov assumptions (1, 2, 3).

Proposition 4.3 (Unbiasedness of OLS)**.** *Under assumptions 1 (linearity: $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$) and 2 (zero conditional mean: $\mathbb{E}[\boldsymbol{\epsilon}|X] = \mathbf{0}$), the OLS estimator $\hat{\boldsymbol{\beta}}$ is conditionally unbiased for $\boldsymbol{\beta}$, given X.*

$$\mathbb{E}[\hat{\boldsymbol{\beta}}|X] = \boldsymbol{\beta}$$

Proof. Substitute $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ into the formula for $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (X^T X)^{-1} X^T Y \\
&= (X^T X)^{-1} X^T (X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= (X^T X)^{-1} X^T X \boldsymbol{\beta} + (X^T X)^{-1} X^T \boldsymbol{\epsilon} \quad \text{(Distributing)} \\
&= I_{p+1}\boldsymbol{\beta} + (X^T X)^{-1} X^T \boldsymbol{\epsilon} \quad \text{(Since } (X^T X)^{-1} X^T X = I\text{)} \\
&= \boldsymbol{\beta} + (X^T X)^{-1} X^T \boldsymbol{\epsilon}
\end{aligned}$$

This shows that the OLS estimator is the true value plus a linear combination of the errors. Now, take the conditional expectation given X. Let $A = (X^T X)^{-1} X^T$. Note that A is constant given X.

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}|X] &= \mathbb{E}[\boldsymbol{\beta} + A\boldsymbol{\epsilon}|X] \\
&= \mathbb{E}[\boldsymbol{\beta}|X] + \mathbb{E}[A\boldsymbol{\epsilon}|X] \quad \text{(Linearity of E)} \\
&= \boldsymbol{\beta} + A\mathbb{E}[\boldsymbol{\epsilon}|X] \quad \text{(Since } \boldsymbol{\beta} \text{ and } A \text{ are constant given } X\text{)} \\
&= \boldsymbol{\beta} + A \cdot \mathbf{0} \quad \text{(Using Assumption 2: } \mathbb{E}[\boldsymbol{\epsilon}|X] = \mathbf{0}\text{)} \\
&= \boldsymbol{\beta}
\end{aligned}$$

Thus, $\hat{\boldsymbol{\beta}}$ is conditionally unbiased for $\boldsymbol{\beta}$. This relies crucially on the model being linear and the errors having zero mean conditional on X. □

Proposition 4.4 (Variance-Covariance Matrix of OLS)**.** *Under assumptions 1, 2, and 3 (linearity, zero conditional mean, homoscedasticity and no correlation:* $\text{Var}(\boldsymbol{\epsilon}|X) = \sigma^2 I_n$*), the conditional variance-covariance matrix of* $\hat{\boldsymbol{\beta}}$*, given* X*, is:*

$$\text{Var}(\hat{\boldsymbol{\beta}}|X) = \sigma^2(X^TX)^{-1}$$

Proof. From the unbiasedness proof, we have $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + A\boldsymbol{\epsilon}$, where $A = (X^TX)^{-1}X^T$. Since $\boldsymbol{\beta}$ is a constant vector, $\text{Var}(\hat{\boldsymbol{\beta}}|X) = \text{Var}(\boldsymbol{\beta} + A\boldsymbol{\epsilon}|X) = \text{Var}(A\boldsymbol{\epsilon}|X)$. Using the property $\text{Var}(MZ) = M\text{Var}(Z)M^T$ for a constant matrix M (from Prop 1.7), with $M = A$:

$$\text{Var}(\hat{\boldsymbol{\beta}}|X) = A\text{Var}(\boldsymbol{\epsilon}|X)A^T$$

Now, use Assumption 3: $\text{Var}(\boldsymbol{\epsilon}|X) = \sigma^2 I_n$.

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}|X) &= A(\sigma^2 I_n)A^T \\
&= \sigma^2 A I_n A^T \\
&= \sigma^2 A A^T \\
&= \sigma^2 \left[(X^TX)^{-1}X^T\right] \left[(X^TX)^{-1}X^T\right]^T \quad \text{(Substituting A)} \\
&= \sigma^2 (X^TX)^{-1}X^T(X^T)^T((X^TX)^{-1})^T \quad \text{(Using } (CD)^T = D^TC^T\text{)} \\
&= \sigma^2 (X^TX)^{-1}X^TX(X^TX)^{-1} \quad \text{(Since } (A^T)^T = A \text{ and } (X^TX)^{-1} \text{ is symmetric)} \\
&= \sigma^2 (X^TX)^{-1}(X^TX)(X^TX)^{-1} \\
&= \sigma^2 (X^TX)^{-1} I_{p+1} \\
&= \sigma^2 (X^TX)^{-1}
\end{aligned}$$

This derivation relies on all three assumptions: linearity (to write $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + A\boldsymbol{\epsilon}$), zero conditional mean (implicitly used in the variance definition relative to the mean $\boldsymbol{\beta}$), and crucially, homoscedasticity and no correlation (to substitute $\text{Var}(\boldsymbol{\epsilon}|X) = \sigma^2 I$). If errors were heteroscedastic or correlated, $\text{Var}(\boldsymbol{\epsilon}|X)$ would be a different matrix Ω, and the result would be $\text{Var}(\hat{\boldsymbol{\beta}}|X) = A\Omega A^T$, which is more complex (leading to Generalized Least Squares). □

4.2 Residuals and Projection Matrices

Let's revisit residuals and their connection to projection matrices.

- Fitted values: $\hat{Y} = X\hat{\boldsymbol{\beta}} = X(X^TX)^{-1}X^TY = P_XY$. Recall $P_X = X(X^TX)^{-1}X^T$ is the projection matrix onto the column space of X, $Col(X)$. It's symmetric ($P_X^T = P_X$) and idempotent ($P_XP_X = P_X$).
- Residual vector: $e = Y - \hat{Y} = Y - P_XY = (I - P_X)Y = M_XY$. Recall $M_X = I - P_X$ is the projection matrix onto the space orthogonal to $Col(X)$. It's also symmetric and idempotent, and $M_XP_X = P_XM_X = \mathbf{0}$ (the zero matrix).

Proposition 4.5 (Properties of Residuals)**.** *1.* ***Mean of Residuals (Mathematical Property):*** *If the model includes an intercept (i.e., the first column of* X *is* $\mathbf{1}$*, the vector of ones), then the sum of the OLS residuals is always exactly zero:* $\sum_{i=1}^n e_i = \mathbf{1}^Te = 0$.

2. ***Expectation of Residuals (Statistical Property):*** *Under assumptions 1 and 2 ($\mathbb{E}[Y|X] = X\boldsymbol{\beta}$), $\mathbb{E}[e|X] = \mathbf{0}$.*

Proof. 1. $\mathbf{1}^T e = \mathbf{1}^T M_X Y$. Since M_X is symmetric ($M_X^T = M_X$), this is $(M_X \mathbf{1})^T Y$. If $\mathbf{1}$ is a column in X, then $\mathbf{1} \in Col(X)$. The projection $M_X \mathbf{1}$ projects $\mathbf{1}$ onto the space orthogonal to $Col(X)$. Since $\mathbf{1}$ is already *in* $Col(X)$, this projection must be zero. So $M_X \mathbf{1} = \mathbf{0}$. Thus $\mathbf{1}^T e = \mathbf{0}^T Y = 0$. This is purely algebraic, relying only on the presence of an intercept (which puts $\mathbf{1}$ in $Col(X)$).

2. $\mathbb{E}[e|X] = \mathbb{E}[M_X Y|X]$. Since M_X is constant given X, $\mathbb{E}[e|X] = M_X \mathbb{E}[Y|X]$. Using assumptions 1 and 2, we know $\mathbb{E}[Y|X] = X\boldsymbol{\beta}$. So, $\mathbb{E}[e|X] = M_X(X\boldsymbol{\beta})$. Since the columns of $X\boldsymbol{\beta}$ are linear combinations of columns of X, $X\boldsymbol{\beta} \in Col(X)$. Therefore, the projection onto the orthogonal complement is zero: $M_X(X\boldsymbol{\beta}) = \mathbf{0}$. This relies on the model assumptions yielding $\mathbb{E}[Y|X] = X\boldsymbol{\beta}$.

So, the fact that $\sum e_i = 0$ (when there's an intercept) is always true by construction of OLS. The fact that $\mathbb{E}[e|X] = \mathbf{0}$ (meaning the errors average to zero in expectation, even conditional on X) requires the model assumptions to hold. □

Lemma 4.6 (Projection Property for Nested Models)**.** *Let $L \subset M$ be two vector subspaces of $\mathbb{R}^n$ (e.g., $L = Col(X_L)$ and $M = Col(X_M)$ where X_L is a subset of the columns of X_M). Let P_L and P_M be the orthogonal projection matrices onto L and M respectively. Then $P_L P_M = P_M P_L = P_L$.*

Proof. Proof that $P_M P_L = P_L$: Take any vector $v \in \mathbb{R}^n$. $P_L v$ is the projection of v onto L. By definition, $P_L v \in L$. Since $L \subset M$, it follows that $P_L v$ is also in M. Projecting a vector that is already in M onto M leaves it unchanged. Therefore, $P_M(P_L v) = P_L v$. Since this holds for all v, we have $P_M P_L = P_L$.

Proof that $P_L P_M = P_L$: Since P_L and P_M are projection matrices, they are symmetric ($P_L^T = P_L$, $P_M^T = P_M$). Using the result we just proved ($P_M P_L = P_L$) and taking the transpose: $(P_M P_L)^T = P_L^T \ P_L^T P_M^T = P_L^T \ P_L P_M = P_L$. Thus, both products equal P_L. □

Example 4.7 (Fitted Values in Nested Models)**.** Suppose we have a ”full” model estimated by OLS: $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with fitted values $\hat{Y} = P_X Y$. Now consider a ”reduced” model using only a subset of the columns of X, say X_L, where $Col(X_L) \subset Col(X)$. The OLS fit for this reduced model is $Y = X_L \hat{\boldsymbol{\gamma}} + e_L$, with fitted values $\hat{Y}_L = P_L Y$, where P_L projects onto $Col(X_L)$.

Consider applying the projection P_L to the fitted values from the full model, $\hat{Y}$:

$$P_L \hat{Y} = P_L(P_X Y)$$

Since $Col(X_L) \subset Col(X)$, we can apply Lemma 4.6 with $L = Col(X_L)$ and $M = Col(X)$. The lemma states $P_L P_X = P_L$. Substituting this into the equation:

$$P_L \hat{Y} = (P_L P_X) Y = P_L Y$$

But we know that $P_L Y$ are the fitted values from the reduced model, $\hat{Y}_L$. Therefore,

$$P_L \hat{Y} = \hat{Y}_L$$

This means if you take the predictions from a larger model and orthogonally project them onto the subspace defined by a smaller (nested) model, you obtain exactly the predictions from fitting the smaller model directly. This geometric insight is fundamental to understanding analysis of variance (ANOVA) tables and tests for comparing nested linear models.

This concludes our review and extension of random vectors and our introduction to the linear model framework and the OLS estimator. Next time, we will build upon this foundation, exploring inference and diagnostics.