Regression and Statistical Models

Tutorial 4: Multivariate Distributions and the Linear Model

Undergraduate Mathematics Educator

October 2023 (Based on Prior Lecture Notes)

Abstract

Welcome! These notes cover fundamental concepts related to multivariate random variables, focusing on expectations and covariance matrices. We will then transition to the cornerstone of this course: the linear model. We'll explore its matrix formulation, underlying assumptions, and key properties derived from those assumptions. We will work through several examples to solidify understanding, including some problems adapted from past exams. Pay close attention to the distinctions between assumptions and derived results, and how different modeling scenarios align with the standard linear model framework.

1 Multivariate Random Variables: Expectations and Covariance

Often in statistics, we deal not just with single random variables, but with collections of them. Understanding their joint behavior is crucial. Let's start by defining some key concepts for random vectors.

Definition 1.1 (Random Vector)**.** A vector $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_n)^T$ whose components Z_i are random variables is called a *random vector*.

Definition 1.2 (Expectation of a Random Vector/Matrix)**.** Let $\boldsymbol{Z} = (Z_1, \dots, Z_n)^T$ be a random vector. Its *expectation* is the vector of the expectations of its components:

$$\mathbb{E}[\boldsymbol{Z}] = \begin{pmatrix} \mathbb{E}[Z_1] \\ \vdots \\ \mathbb{E}[Z_n] \end{pmatrix}$$

Similarly, if $\boldsymbol{A}$ is a random matrix with entries A_{ij}, its expectation is the matrix of the expectations of its entries: $(\mathbb{E}[\boldsymbol{A}])_{ij} = \mathbb{E}[A_{ij}]$.

Expectation behaves linearly, even with vectors and matrices. Let $\boldsymbol{Z}, \boldsymbol{W}$ be random vectors, $\boldsymbol{A}, \boldsymbol{B}$ be fixed (non-random) matrices of appropriate dimensions, and $\boldsymbol{C}$ be a fixed vector.

Proposition 1.3 (Properties of Expectation)**.** *1.* $\mathbb{E}[\boldsymbol{Z} + \boldsymbol{W}] = \mathbb{E}[\boldsymbol{Z}] + \mathbb{E}[\boldsymbol{W}]$

2. $\mathbb{E}[\boldsymbol{AZB}] = \boldsymbol{A}\,\mathbb{E}[\boldsymbol{Z}]\boldsymbol{B}$ *(Note: $\boldsymbol{B}$ must be 1×1 or $\boldsymbol{Z}$ must be a matrix for this to be generally applicable. If $\boldsymbol{Z}$ is a vector, often we see $\mathbb{E}[\boldsymbol{AZ}] = \boldsymbol{A}\,\mathbb{E}[\boldsymbol{Z}]$ or $\mathbb{E}[\boldsymbol{Z}^T\boldsymbol{B}] = \mathbb{E}[\boldsymbol{Z}]^T\boldsymbol{B}$.)*

3. $\mathbb{E}[\boldsymbol{AZ} + \boldsymbol{C}] = \boldsymbol{A}\,\mathbb{E}[\boldsymbol{Z}] + \boldsymbol{C}$ *(Follows from 1 and 2)*

While expectation describes the "center" of the distribution, the covariance matrix describes the spread and linear relationships between components.

Definition 1.4 (Covariance and Variance-Covariance Matrix)**.** Let $\boldsymbol{Z} \in \mathbb{R}^n$ and $\boldsymbol{W} \in \mathbb{R}^m$ be random vectors.

1. The *covariance matrix* between $\boldsymbol{Z}$ and $\boldsymbol{W}$ is the $n \times m$ matrix:

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

   The (i, j)-th entry is $\mathrm{Cov}(Z_i, W_j)$.

2. The *variance-covariance matrix* (or simply variance matrix) of $\boldsymbol{Z}$ is the $n \times n$ matrix:

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

Proposition 1.5. *The (i, j)-th entry of the variance-covariance matrix* $\mathrm{Var}(\boldsymbol{Z})$ *is the covariance between Z_i and Z_j:*

$$(\mathrm{Var}(\boldsymbol{Z}))_{ij} = \mathrm{Cov}(Z_i, Z_j)$$

Consequently, $\mathrm{Var}(\boldsymbol{Z})$ *is a symmetric matrix, i.e.,* $(\mathrm{Var}(\boldsymbol{Z}))_{ij} = (\mathrm{Var}(\boldsymbol{Z}))_{ji}$.

Proof. Let $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Z}]$. The (i, j)-th entry of $\mathrm{Var}(\boldsymbol{Z})$ is given by:

$$\begin{aligned}
(\mathrm{Var}(\boldsymbol{Z}))_{ij} &= \left(\mathbb{E}\left[(\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})^T\right]\right)_{ij} \\
&= \mathbb{E}\left[((\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})^T)_{ij}\right] \quad \text{(Expectation is element-wise)} \\
&= \mathbb{E}\left[(\boldsymbol{Z} - \boldsymbol{\mu})_i(\boldsymbol{Z} - \boldsymbol{\mu})_j\right] \quad \text{(Definition of matrix multiplication)} \\
&= \mathbb{E}\left[(Z_i - \mathbb{E}[Z_i])(Z_j - \mathbb{E}[Z_j])\right] \\
&= \mathrm{Cov}(Z_i, Z_j)
\end{aligned}$$

Since $\mathrm{Cov}(Z_i, Z_j) = \mathrm{Cov}(Z_j, Z_i)$, we have $(\mathrm{Var}(\boldsymbol{Z}))_{ij} = (\mathrm{Var}(\boldsymbol{Z}))_{ji}$. □

Proposition 1.6 (Properties of Covariance Matrices)**.** *Let $\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{R}$ be random vectors, $\boldsymbol{A}, \boldsymbol{B}$ be fixed matrices of appropriate dimensions, and $\boldsymbol{a}$ be a fixed vector. The following properties hold:*

1. $\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{W}) = \mathrm{Cov}(\boldsymbol{W}, \boldsymbol{Z})^\top$
2. $\mathrm{Cov}(\boldsymbol{Z} + \boldsymbol{R}, \boldsymbol{W}) = \mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{W}) + \mathrm{Cov}(\boldsymbol{R}, \boldsymbol{W})$
3. $\mathrm{Cov}(\boldsymbol{A}\boldsymbol{Z}, \boldsymbol{B}\boldsymbol{W}) = \boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{W})\boldsymbol{B}^\top$
4. $\mathrm{Var}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\,\mathrm{Var}(\boldsymbol{Z})\boldsymbol{A}^\top$ *(From property 3 with $\boldsymbol{W} = \boldsymbol{Z}, \boldsymbol{B} = \boldsymbol{A}$)*
5. $\mathrm{Var}(\boldsymbol{a}^\top\boldsymbol{Z}) = \boldsymbol{a}^\top\,\mathrm{Var}(\boldsymbol{Z})\boldsymbol{a}$ *(From property 4, treating $\boldsymbol{a}^T$ as a $1 \times n$ matrix A)*
6. $\mathrm{Var}(\boldsymbol{Z})$ *is a symmetric and positive semidefinite matrix. (From property 5:* $\mathrm{Var}(\boldsymbol{a}^\top\boldsymbol{Z}) \geq 0$ *for any $\boldsymbol{a}$, which means* $\boldsymbol{a}^\top\,\mathrm{Var}(\boldsymbol{Z})\boldsymbol{a} \geq 0$*.)*

Remark 1.7. You will be asked to prove these properties in the upcoming homework assignment. They are fundamental tools for manipulating variances and covariances of linear combinations of random variables.

Example 1.8 (Multivariate Normal Transformation)**.** Let $Z_1, \ldots, Z_5$ be independent and identically distributed (iid) random variables following the standard normal distribution, $Z_i \sim \mathcal{N}(0, 1)$.

(a) Find the expectation vector and variance-covariance matrix of the random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_5)^T$.

Solution: Since each $Z_i \sim \mathcal{N}(0, 1)$, we have $\mathbb{E}[Z_i] = 0$ and $\mathrm{Var}(Z_i) = 1$. Because the Z_i are independent, $\mathrm{Cov}(Z_i, Z_j) = 0$ for $i \neq j$. The expectation vector is:

$$\mathbb{E}[\boldsymbol{Z}] = (\mathbb{E}[Z_1], \ldots, \mathbb{E}[Z_5])^T = (0, \ldots, 0)^T = \boldsymbol{0}_5$$

The variance-covariance matrix is:

$$\mathrm{Var}(\boldsymbol{Z})_{ij} = \mathrm{Cov}(Z_i, Z_j) = \begin{cases} \mathrm{Var}(Z_i) = 1 & \text{if } i = j \\ \mathrm{Cov}(Z_i, Z_j) = 0 & \text{if } i \neq j \end{cases}$$

So, $\mathrm{Var}(\boldsymbol{Z})$ is the 5×5 identity matrix:

$$\mathrm{Var}(\boldsymbol{Z}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \boldsymbol{I}_5$$

(b) Define the transformation $A : \mathbb{R}^5 \to \mathbb{R}^3$ by the matrix:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Is this transformation linear? Calculate the expectation vector and variance-covariance matrix of the transformed vector $\boldsymbol{W} = A\boldsymbol{Z}$.

Solution: Yes, the transformation $\boldsymbol{Z} \mapsto \boldsymbol{AZ}$ is a linear transformation because it is defined by matrix multiplication. Using the properties of expectation and covariance:

$$\mathbb{E}[\boldsymbol{W}] = \mathbb{E}[\boldsymbol{AZ}] = \boldsymbol{A}\,\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{A}\boldsymbol{0}_5 = \boldsymbol{0}_3$$

The expectation of the transformed vector is the zero vector in $\mathbb{R}^3$.

$$\mathrm{Var}(\boldsymbol{W}) = \mathrm{Var}(\boldsymbol{AZ}) = \boldsymbol{A}\,\mathrm{Var}(\boldsymbol{Z})\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{I}_5\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{A}^T$$

Let's compute $\boldsymbol{AA}^T$:

$$\begin{aligned} \boldsymbol{AA}^T &= \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} (1)(1) + (1)(1) + 0 + 0 + 0 & (1)(0) + (1)(0) + 0 + 0 + 0 & (1)(0) + (1)(0) + 0 + 0 + 0 \\ 0 + 0 + (1)(0) + (1)(0) + 0 & 0 + 0 + (1)(1) + (1)(1) + 0 & 0 + 0 + (1)(0) + (1)(0) + 0 \\ 0 + 0 + 0 + 0 + (1)(0) & 0 + 0 + 0 + 0 + (1)(0) & 0 + 0 + 0 + 0 + (1)(1) \end{pmatrix} \\ &= \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

So, $\mathrm{Var}(\boldsymbol{W}) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Notice that the components of $\boldsymbol{W} = (Z_1 + Z_2, Z_3 + Z_4, Z_5)^T$ are uncorrelated, but W_1 and W_2 have variance 2, while W_3 has variance 1.

(c) Would your answers to (a) and (b) change if it were only known that $Z_1, \ldots, Z_5$ were sampled such that $\mathbb{E}[Z_i] = 0$, $\mathrm{Var}(Z_i) = 1$ for all i, and $\mathrm{Cov}(Z_i, Z_j) = 0$ for $i \neq j$ (i.e., they are uncorrelated with mean 0 and variance 1, but not necessarily normally distributed or independent)?

Solution: No, the answers for $\mathbb{E}[\boldsymbol{Z}]$, $\mathrm{Var}(\boldsymbol{Z})$, $\mathbb{E}[\boldsymbol{W}]$, and $\mathrm{Var}(\boldsymbol{W})$ would *not* change. The calculations for the expectation vector and variance-covariance matrix only depend on the first moments ($\mathbb{E}[Z_i]$) and second moments ($\mathbb{E}[Z_i Z_j]$, which determine variances and covariances). They do not depend on the full distributional shape (like normality). The properties $\mathbb{E}[\boldsymbol{A}\boldsymbol{Z}] = \boldsymbol{A}\,\mathbb{E}[\boldsymbol{Z}]$ and $\text{Var}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\,\text{Var}(\boldsymbol{Z})\boldsymbol{A}^T$ hold regardless of the underlying distribution, as long as the expectations and variances exist.

Proposition 1.9 (Variance Ordering and Positive Semidefiniteness)**.** *Let $\boldsymbol{Z}, \boldsymbol{W}$ be random vectors in $\mathbb{R}^p$. The following are equivalent:*

1. *For all constant vectors $\boldsymbol{v} \in \mathbb{R}^p$, $\text{Var}(\boldsymbol{v}^T\boldsymbol{Z}) \geq \text{Var}(\boldsymbol{v}^T\boldsymbol{W})$.*
2. *The matrix $\boldsymbol{B} := \text{Var}(\boldsymbol{Z}) - \text{Var}(\boldsymbol{W})$ is positive semidefinite.*

(Recall that a symmetric matrix $\boldsymbol{B}$ is positive semidefinite if $\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{v} \geq 0$ for all vectors $\boldsymbol{v}$.)

Proof. Using property 5 of covariance matrices:

$$\text{Var}(\boldsymbol{v}^T\boldsymbol{Z}) = \boldsymbol{v}^T\,\text{Var}(\boldsymbol{Z})\boldsymbol{v}$$

$$\text{Var}(\boldsymbol{v}^T\boldsymbol{W}) = \boldsymbol{v}^T\,\text{Var}(\boldsymbol{W})\boldsymbol{v}$$

Therefore, statement (1) is equivalent to:

$$\boldsymbol{v}^T\,\text{Var}(\boldsymbol{Z})\boldsymbol{v} \geq \boldsymbol{v}^T\,\text{Var}(\boldsymbol{W})\boldsymbol{v} \quad \text{for all } \boldsymbol{v} \in \mathbb{R}^p$$

Rearranging gives:

$$\boldsymbol{v}^T(\text{Var}(\boldsymbol{Z}) - \text{Var}(\boldsymbol{W}))\boldsymbol{v} \geq 0 \quad \text{for all } \boldsymbol{v} \in \mathbb{R}^p$$

This is precisely the definition of the matrix $\boldsymbol{B} = \text{Var}(\boldsymbol{Z}) - \text{Var}(\boldsymbol{W})$ being positive semidefinite. Note that $\boldsymbol{B}$ is symmetric since $\text{Var}(\boldsymbol{Z})$ and $\text{Var}(\boldsymbol{W})$ are symmetric. □

Remark 1.10. The condition that $\text{Var}(\boldsymbol{Z}) - \text{Var}(\boldsymbol{W})$ is positive semidefinite provides a way to compare the "overall dispersion" of two random vectors, often denoted as $\text{Var}(\boldsymbol{Z}) \geq \text{Var}(\boldsymbol{W})$ in the Loewner ordering of matrices. This concept is important in comparing the efficiency of estimators, for example. The source mentioned a condition involving $B = C^TC$ or $B = C^2$; this is related because a symmetric matrix is positive semidefinite if and only if it can be expressed as C^TC for some matrix C (e.g., via Cholesky decomposition or spectral decomposition).

2 The Linear Model

We now shift focus to the primary topic of the course: the linear model. This model forms the basis for regression analysis and many other statistical techniques.

2.1 Model Definition and Matrix Notation

Suppose we have n observations. For each observation $i = 1, \ldots, n$, we have a response variable Y_i and a set of p predictor variables (or features) $X_{i1}, \ldots, X_{ip}$.

The *linear model* posits that the relationship between the response and the predictors is approximately linear, incorporating some random error:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

Here:

- Y_i is the response variable for the i-th observation.
- X_{ij} is the value of the j-th predictor for the i-th observation.

- β_0 is the intercept term (the expected value of Y when all X_j are zero).
- $\beta_1, \ldots, \beta_p$ are the coefficients associated with each predictor, representing the change in Y for a one-unit change in the corresponding X_j, holding other predictors constant. These are unknown parameters we typically want to estimate.
- ϵ_i is the random error term for the i-th observation, representing variability in Y_i not explained by the predictors.

It's incredibly convenient to express this model using matrix notation. Let's define:

- The response vector $\boldsymbol{Y} \in \mathbb{R}^n$: $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$
- The design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$:
$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}$$
(Note the first column of ones, corresponding to the intercept β_0).
- The parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$: $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$
- The error vector $\boldsymbol{\epsilon} \in \mathbb{R}^n$: $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$

With these definitions, the entire set of n equations can be written compactly as:

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

2.2 Standard Assumptions

The utility of the linear model comes from making certain assumptions about the error terms ϵ_i. The standard assumptions are:

1. **Linearity:** The model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ correctly describes the relationship. This implies $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}$.
2. **Zero Mean Error:** The expected value of the errors is zero: $\mathbb{E}[\epsilon_i] = 0$ for all i, or in matrix form, $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.
3. **Homoscedasticity:** The variance of the errors is constant for all observations: $\mathrm{Var}(\epsilon_i) = \sigma^2$ for all i, where σ^2 is some positive constant.
4. **Uncorrelated Errors:** The errors for different observations are uncorrelated: $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$.

Assumptions 2, 3, and 4 can be concisely written using the variance-covariance matrix of the error vector:

$$\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0} \quad \text{and} \quad \text{Cov}(\boldsymbol{\epsilon}) = \text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

An additional assumption often made, particularly for inference (hypothesis testing, confidence intervals), is:

5. **Normality:** The errors are normally distributed: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. This implies $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$.

This is referred to as the *Normal Linear Model.*

2.3 Estimation and Key Quantities

Our goal is often to estimate the unknown parameter vector $\boldsymbol{\beta}$. The most common method is *Ordinary Least Squares (OLS).* The OLS estimator $\hat{\boldsymbol{\beta}}$ is the vector that minimizes the sum of squared differences between the observed responses Y_i and the values predicted by the linear function of X_i.

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{b} \in \mathbb{R}^{p+1}} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}||^2 = \text{argmin}_{\boldsymbol{b}} \sum_{i=1}^{n} (Y_i - (\boldsymbol{X}\boldsymbol{b})_i)^2$$

Under standard conditions (specifically, that $\boldsymbol{X}$ has full column rank, meaning its columns are linearly independent), the OLS estimator has a closed-form solution:

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

Once we have $\hat{\boldsymbol{\beta}}$, we can define:

- **Fitted values:** The predicted values on the regression line/plane:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

- **Residuals:** The differences between observed and fitted values:

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$$

  The residuals $\boldsymbol{e}$ are our empirical estimate of the unobservable errors $\boldsymbol{\epsilon}$.

Remark 2.1. Unless stated otherwise, $\hat{\boldsymbol{\beta}}$ will refer to the OLS estimator.

Example 2.2 (Assumptions vs. Results)**.** Consider the following statements related to the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Classify each as either a fundamental assumption of the model (or OLS procedure) or a mathematical result derived from the assumptions/definitions. Assume $\boldsymbol{X}$ is fixed (non-random).

1. $\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{b}} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}||^2$: **Definition/Result.** This is the definition of the OLS estimator $\hat{\boldsymbol{\beta}}$. It's the result of applying the least squares principle.

2. $\boldsymbol{X}\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]$ (or $\mathbb{E}[\boldsymbol{Y}]$ if $\boldsymbol{X}$ is fixed): **Assumption.** This is the core linearity assumption. It states that the conditional expectation of the response is a linear function of the predictors.

3. $\mathbb{E}[e_i] = 0$: **Result.** The OLS residuals have zero mean, $\mathbb{E}[\boldsymbol{e}] = \mathbb{E}[\boldsymbol{Y} - \hat{\boldsymbol{Y}}] = \mathbb{E}[\boldsymbol{Y}] - \mathbb{E}[\boldsymbol{X}\hat{\boldsymbol{\beta}}]$. Assuming $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$ and that $\hat{\boldsymbol{\beta}}$ is unbiased (which we will show later), $\mathbb{E}[\boldsymbol{e}] = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\,\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$. More directly, $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{P}_{\mathcal{C}})\boldsymbol{Y}$ where $\boldsymbol{P}_{\mathcal{C}}$ is the projection onto the column space of $\boldsymbol{X}$. $\mathbb{E}[\boldsymbol{e}] = (\boldsymbol{I} - \boldsymbol{P}_{\mathcal{C}})\,\mathbb{E}[\boldsymbol{Y}] = (\boldsymbol{I} - \boldsymbol{P}_{\mathcal{C}})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{P}_{\mathcal{C}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$ (since $\boldsymbol{X}\boldsymbol{\beta}$ is in the column space of $\boldsymbol{X}$, $P_{\mathcal{C}}$ leaves it unchanged).

4. $\mathbb{E}[\epsilon_i] = 0$: **Assumption.** This is the standard assumption of zero mean errors.

5. $\boldsymbol{X}^T\boldsymbol{e} = \boldsymbol{0}$: **Result.** This is a direct consequence of the OLS normal equations. $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Then $\boldsymbol{X}^T\boldsymbol{e} = \boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{0}$. Geometrically, it means the residual vector is orthogonal to the column space of $\boldsymbol{X}$.

6. $\text{Var}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$: **Result (under fixed X).** If we treat $\boldsymbol{X}$ as fixed (non-random), then $\text{Var}(\boldsymbol{Y}) = \text{Var}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \text{Var}(\boldsymbol{\epsilon})$. If we assume $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$, then $\text{Var}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$. If $\boldsymbol{X}$ is random, this is generally not true.

Example 2.3 (Scenarios and Model Assumptions)**.** Consider the following scenarios describing how data (X_i, Y_i) might arise. For each, specify the distributions (or nature) of X_i, Y_i, ϵ_i and $Y_i|X_i$. Identify which assumptions of the standard linear model ($\mathbb{E}[\epsilon_i|X_i] = 0$, $\text{Var}(\epsilon_i|X_i) = \sigma^2$, errors uncorrelated, linearity) hold. Let ϵ_i generally be iid $\mathcal{N}(0, \sigma^2)$ and independent of X_i unless stated otherwise.

1. **Fixed Design:** $X_i \in \mathbb{R}^p$ are predetermined, fixed constants. $Y_i = X_i^T\boldsymbol{\beta} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ iid.

   - X_i: Fixed vectors.
   - ϵ_i: iid $\mathcal{N}(0, \sigma^2)$.
   - $Y_i|X_i$: Since X_i is fixed, this is just the distribution of Y_i. $Y_i \sim \mathcal{N}(X_i^T\boldsymbol{\beta}, \sigma^2)$. Y_i are independent.
   - Assumptions: Linearity holds ($\mathbb{E}[Y_i|X_i] = X_i^T\boldsymbol{\beta}$), $\mathbb{E}[\epsilon_i|X_i] = \mathbb{E}[\epsilon_i] = 0$. $\text{Var}(\epsilon_i|X_i) = \text{Var}(\epsilon_i) = \sigma^2$ (Homoscedasticity). Errors are uncorrelated (actually independent). Normality holds. All standard assumptions are met.

2. **Random Design (Normal):** $X_i \in \mathbb{R}^p$ are iid random vectors, e.g., $X_i \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$. $Y_i = X_i^T\boldsymbol{\beta} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ iid, and ϵ_i are independent of X_i.

   - X_i: iid random vectors (Normal).
   - ϵ_i: iid $\mathcal{N}(0, \sigma^2)$, independent of X_i.
   - $Y_i|X_i$: Conditional on X_i, $Y_i \sim \mathcal{N}(X_i^T\boldsymbol{\beta}, \sigma^2)$.
   - Y_i: Random variable (its distribution is a mixture, often normal if X_i is normal). Y_i are generally dependent unless $\boldsymbol{\beta} = 0$.
   - Assumptions: Conditional linearity $\mathbb{E}[Y_i|X_i] = X_i^T\boldsymbol{\beta}$ holds. $\mathbb{E}[\epsilon_i|X_i] = \mathbb{E}[\epsilon_i] = 0$ (due to independence). $\text{Var}(\epsilon_i|X_i) = \text{Var}(\epsilon_i) = \sigma^2$ (Homoscedasticity, due to independence). Errors ϵ_i are uncorrelated/independent. Normality of errors holds. All standard *conditional* assumptions hold. The properties of OLS estimators often rely on these conditional assumptions.

3. **Random Design (Uniform):** $X_i \in \mathbb{R}^1$ are iid $U(-1, 1)$. $Y_i = X_i\beta + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ iid, independent of X_i.

   - X_i: iid $U(-1, 1)$.

- ϵ_i: iid $\mathcal{N}(0, \sigma^2)$, independent of X_i.
- $Y_i|X_i$: Conditional on X_i, $Y_i \sim \mathcal{N}(X_i\beta, \sigma^2)$.
- Y_i: Random variable (non-normal distribution).
- Assumptions: Conditional linearity $\mathbb{E}[Y_i|X_i] = X_i\beta$ holds. $\mathbb{E}[\epsilon_i|X_i] = 0$. $\text{Var}(\epsilon_i|X_i) = \sigma^2$ (Homoscedasticity). Errors are uncorrelated/independent. Normality of errors holds. Standard conditional assumptions are met, but Y_i itself is not normal.

4. **Non-linear Relationship:** $X_i \in \mathbb{R}^1$ are iid $\mathcal{N}(0, 1)$. $Y_i = X_i^2\beta + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ iid, independent of X_i.

   - X_i: iid $\mathcal{N}(0, 1)$.
   - ϵ_i: iid $\mathcal{N}(0, \sigma^2)$, independent of X_i.
   - $Y_i|X_i$: Conditional on X_i, $Y_i \sim \mathcal{N}(X_i^2\beta, \sigma^2)$.
   - Y_i: Random variable.
   - Assumptions: The relationship $\mathbb{E}[Y_i|X_i] = X_i^2\beta$ is *not* linear in X_i. The linearity assumption fails if we model Y_i vs X_i. However, if we define a new predictor $Z_i = X_i^2$, then $Y_i = Z_i\beta + \epsilon_i$, and the model *is* linear in Z_i. $\mathbb{E}[\epsilon_i|X_i] = 0$ holds. $\text{Var}(\epsilon_i|X_i) = \sigma^2$ holds. Errors are uncorrelated/independent. Normality of errors holds. The standard assumptions hold for the model Y_i vs $Z_i = X_i^2$, but not for Y_i vs X_i.

5. **Panel Data Structure:** $X_{it} \in \mathbb{R}^p$ are fixed design variables for individual i at time t. $Y_{it} = X_{it}^T\boldsymbol{\beta} + \epsilon_{it}$, where $\epsilon_{it} \sim \mathcal{N}(0, \sigma^2)$ iid across both i and t.

   - X_{it}: Fixed vectors.
   - ϵ_{it}: iid $\mathcal{N}(0, \sigma^2)$.
   - $Y_{it}|X_{it}$: $Y_{it} \sim \mathcal{N}(X_{it}^T\boldsymbol{\beta}, \sigma^2)$.
   - Assumptions: If we stack all Y_{it} into a single vector $\boldsymbol{Y}$ and all X_{it}^T into a large design matrix $\boldsymbol{X}$, the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ holds. Linearity holds. $\mathbb{E}[\epsilon_{it}] = 0$. $\text{Var}(\epsilon_{it}) = \sigma^2$ (Homoscedasticity). Errors are uncorrelated (by assumption). Normality holds. All standard assumptions are met. (Note: In practice, panel data often has correlated errors within individuals, requiring more advanced models).

3 Properties of Estimators and Related Quantities

Let's delve into some properties related to common statistics used in linear models and basic inference.

Example 3.1 (Orthogonal Projection Matrix)**.** Let $\boldsymbol{v} \in \mathbb{R}^n$ be a non-zero vector, $\boldsymbol{v} \neq \boldsymbol{0}$. Show that the matrix $\boldsymbol{P} = \frac{\boldsymbol{v}\boldsymbol{v}^T}{||\boldsymbol{v}||^2}$ is an orthogonal projection matrix. What is the rank of this matrix?

Solution: An orthogonal projection matrix must be symmetric ($\boldsymbol{P} = \boldsymbol{P}^T$) and idempotent ($\boldsymbol{P}^2 = \boldsymbol{P}$).

1. **Symmetry:** We need to show $\boldsymbol{P}^T = \boldsymbol{P}$.

$$\boldsymbol{P}^T = \left(\frac{\boldsymbol{v}\boldsymbol{v}^T}{||\boldsymbol{v}||^2}\right)^T = \frac{(\boldsymbol{v}\boldsymbol{v}^T)^T}{||\boldsymbol{v}||^2} = \frac{(\boldsymbol{v}^T)^T\boldsymbol{v}^T}{||\boldsymbol{v}||^2} = \frac{\boldsymbol{v}\boldsymbol{v}^T}{||\boldsymbol{v}||^2} = \boldsymbol{P}$$

So, $\boldsymbol{P}$ is symmetric.

2. **Idempotence:** We need to show $\boldsymbol{P}^2 = \boldsymbol{P}$.

$$\begin{aligned}
\boldsymbol{P}^2 &= \left(\frac{\boldsymbol{v}\boldsymbol{v}^T}{||\boldsymbol{v}||^2}\right)\left(\frac{\boldsymbol{v}\boldsymbol{v}^T}{||\boldsymbol{v}||^2}\right) \\
&= \frac{1}{(||\boldsymbol{v}||^2)^2}(\boldsymbol{v}\boldsymbol{v}^T)(\boldsymbol{v}\boldsymbol{v}^T) \\
&= \frac{1}{||\boldsymbol{v}||^4}\boldsymbol{v}(\boldsymbol{v}^T\boldsymbol{v})\boldsymbol{v}^T \quad \text{(associativity of matrix multiplication)}
\end{aligned}$$

Recognize that $\boldsymbol{v}^T\boldsymbol{v} = \sum_{i=1}^n v_i^2 = ||\boldsymbol{v}||^2$, which is a scalar.

$$\boldsymbol{P}^2 = \frac{1}{||\boldsymbol{v}||^4}\boldsymbol{v}(||\boldsymbol{v}||^2)\boldsymbol{v}^T = \frac{||\boldsymbol{v}||^2}{||\boldsymbol{v}||^4}\boldsymbol{v}\boldsymbol{v}^T = \frac{1}{||\boldsymbol{v}||^2}\boldsymbol{v}\boldsymbol{v}^T = \boldsymbol{P}$$

So, $\boldsymbol{P}$ is idempotent.

Since $\boldsymbol{P}$ is symmetric and idempotent, it is an orthogonal projection matrix. It projects vectors onto the subspace spanned by $\boldsymbol{v}$.

Rank: The matrix $\boldsymbol{v}\boldsymbol{v}^T$ is an outer product of a non-zero vector with itself. Every column of $\boldsymbol{v}\boldsymbol{v}^T$ is a multiple of $\boldsymbol{v}$. Specifically, column j is $v_j\boldsymbol{v}$. Since $\boldsymbol{v} \neq \boldsymbol{0}$, the column space is spanned by the single vector $\boldsymbol{v}$. Therefore, the rank of $\boldsymbol{v}\boldsymbol{v}^T$ is 1. Since $\boldsymbol{P}$ is just a scalar multiple of $\boldsymbol{v}\boldsymbol{v}^T$, its rank is also 1.

$$\text{rank}(\boldsymbol{P}) = 1$$

Example 3.2 (Unbiasedness of Sample Variance)**.** Let $Y_1, \ldots, Y_n$ be iid random variables with mean μ and variance σ^2. Show that the sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$, where $\bar{Y} = \frac{1}{n}\sum Y_i$, is an unbiased estimator for σ^2, i.e., $\mathbb{E}[S_n^2] = \sigma^2$.

Solution: We start by expanding the sum of squares term:

$$\begin{aligned}
\sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (Y_i - \mu + \mu - \bar{Y})^2 \\
&= \sum_{i=1}^n [(Y_i - \mu) - (\bar{Y} - \mu)]^2 \\
&= \sum_{i=1}^n [(Y_i - \mu)^2 - 2(Y_i - \mu)(\bar{Y} - \mu) + (\bar{Y} - \mu)^2] \\
&= \sum_{i=1}^n (Y_i - \mu)^2 - 2(\bar{Y} - \mu)\sum_{i=1}^n (Y_i - \mu) + \sum_{i=1}^n (\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^n (Y_i - \mu)^2 - 2(\bar{Y} - \mu)(n\bar{Y} - n\mu) + n(\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^n (Y_i - \mu)^2 - 2n(\bar{Y} - \mu)^2 + n(\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^n (Y_i - \mu)^2 - n(\bar{Y} - \mu)^2
\end{aligned}$$

Now, we take the expectation:

$$\mathbb{E}\left[\sum_{i=1}^n (Y_i - \bar{Y})^2\right] = \mathbb{E}\left[\sum_{i=1}^n (Y_i - \mu)^2\right] - n\,\mathbb{E}\left[(\bar{Y} - \mu)^2\right]$$

We know that $\mathbb{E}[(Y_i - \mu)^2] = \text{Var}(Y_i) = \sigma^2$. So,

$$\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \mu)^2\right] = \sum_{i=1}^{n}\mathbb{E}[(Y_i - \mu)^2] = \sum_{i=1}^{n}\sigma^2 = n\sigma^2$$

Also, $\mathbb{E}[(\bar{Y} - \mu)^2] = \text{Var}(\bar{Y})$. Since Y_i are iid:

$$\text{Var}(\bar{Y}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(Y_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

Substituting these back:

$$\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = n\sigma^2 - n\left(\frac{\sigma^2}{n}\right) = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

Finally, we find the expectation of S_n^2:

$$\mathbb{E}[S_n^2] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

Thus, S_n^2 is an unbiased estimator for σ^2. The division by $(n-1)$ instead of n is precisely what corrects for the bias introduced by using $\bar{Y}$ (an estimate of μ) instead of μ itself.

Example 3.3 (Distribution of Sample Variance under Normality)**.** Assume now that $Y_1, \ldots, Y_n$ are iid $\mathcal{N}(\mu, \sigma^2)$. Prove the result, previously seen, that

$$\frac{(n-1)S_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi_{n-1}^2$$

where χ_{n-1}^2 denotes the chi-squared distribution with $n-1$ degrees of freedom.

Solution Sketch: This is a standard result often proved using Cochran's Theorem or properties of quadratic forms of normal variables. Here's a conceptual outline:

1. **Standardize:** Let $Z_i = (Y_i - \mu)/\sigma$. Then $Z_1, \ldots, Z_n$ are iid $\mathcal{N}(0,1)$. The vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. 2. **Quadratic Form:** We are interested in the sum of squares $\sum(Y_i - \bar{Y})^2$. Let's express this in terms of $\boldsymbol{Z}$. Note $\bar{Y} = \mu + \sigma\bar{Z}$.

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\sigma Z_i - \sigma\bar{Z})^2 = \sigma^2\sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

So, $\frac{1}{\sigma^2}\sum(Y_i - \bar{Y})^2 = \sum(Z_i - \bar{Z})^2$. 3. **Projection Matrix:** Recall the identity $\sum(Z_i - \bar{Z})^2 = \sum Z_i^2 - n\bar{Z}^2$. This can be written as a quadratic form in $\boldsymbol{Z}$:

$$\sum(Z_i - \bar{Z})^2 = \boldsymbol{Z}^T\boldsymbol{Z} - n(\frac{1}{n}\boldsymbol{1}^T\boldsymbol{Z})^2 = \boldsymbol{Z}^T\boldsymbol{Z} - \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{Z} = \boldsymbol{Z}^T(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{Z}$$

where $\boldsymbol{1}$ is the $n \times 1$ vector of ones. 4. **Idempotent Matrix:** Let $\boldsymbol{M} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$. The term $\frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T = \frac{\boldsymbol{1}\boldsymbol{1}^T}{||\boldsymbol{1}||^2}$ is the projection matrix onto the subspace spanned by $\boldsymbol{1}$ (as seen in Example 3.1, since $||\boldsymbol{1}||^2 = n$). Thus, $\boldsymbol{M}$ is the projection matrix onto the subspace orthogonal to $\boldsymbol{1}$. As a projection matrix, $\boldsymbol{M}$ is symmetric and idempotent ($\boldsymbol{M}^2 = \boldsymbol{M}$). 5. **Rank:** The rank of a projection matrix is the dimension of the subspace it projects onto. The space $\mathbb{R}^n$ can be decomposed into the span of $\boldsymbol{1}$ (dimension 1) and its orthogonal complement (dimension $n-1$). $\boldsymbol{M}$ projects onto this orthogonal complement, so $\text{rank}(\boldsymbol{M}) = n-1$. 6. **Distribution of Quadratic Form:** A key theorem states that if $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ and $\boldsymbol{A}$ is a $k \times k$ symmetric, idempotent matrix with rank r, then the quadratic form $\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z} \sim \chi^2_r$. 7. **Conclusion:** Applying this theorem with $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ and $\boldsymbol{A} = \boldsymbol{M} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, which is symmetric, idempotent, and has rank $n-1$, we conclude that:

$$\frac{(n-1)S_n^2}{\sigma^2} = \sum(Z_i - \bar{Z})^2 = \boldsymbol{Z}^T\boldsymbol{M}\boldsymbol{Z} \sim \chi^2_{n-1}$$

This result is fundamental for constructing confidence intervals and hypothesis tests for σ^2 in the normal setting.

4 Further Properties and Exam Problems

Let's explore a few more useful properties and tackle some problems adapted from previous exams.

Example 4.1 (Expected Squared Norm and Trace)**.** Let $\boldsymbol{Z} \in \mathbb{R}^n$ be a random vector.

1. Show that $\mathbb{E}[||\boldsymbol{Z}||^2] = \text{tr}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T])$.
2. Deduce that if $\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{0}$, then $\mathbb{E}[||\boldsymbol{Z}||^2] = \text{tr}(\text{Var}(\boldsymbol{Z}))$.

Justify each step.

Solution: (1) We start with the definition of the squared Euclidean norm:

$$||\boldsymbol{Z}||^2 = \sum_{i=1}^n Z_i^2$$

Taking the expectation:

$$\mathbb{E}[||\boldsymbol{Z}||^2] = \mathbb{E}\left[\sum_{i=1}^n Z_i^2\right] = \sum_{i=1}^n \mathbb{E}[Z_i^2] \quad \text{(Linearity of Expectation)}$$

Now consider the trace of the matrix $\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T]$. The expectation can be moved inside the trace (as trace is a linear operator and expectation is element-wise):

$$\text{tr}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T]) = \mathbb{E}[\text{tr}(\boldsymbol{Z}\boldsymbol{Z}^T)]$$

The trace of the outer product matrix $\boldsymbol{Z}\boldsymbol{Z}^T$ is the sum of its diagonal elements:

$$\text{tr}(\boldsymbol{Z}\boldsymbol{Z}^T) = \sum_{i=1}^n (\boldsymbol{Z}\boldsymbol{Z}^T)_{ii}$$

The (i,i)-th element of $\boldsymbol{Z}\boldsymbol{Z}^T$ is simply $Z_iZ_i = Z_i^2$. So,

$$\text{tr}(\boldsymbol{Z}\boldsymbol{Z}^T) = \sum_{i=1}^n Z_i^2$$

Taking the expectation:

$$\mathbb{E}[\text{tr}(\boldsymbol{Z}\boldsymbol{Z}^T)] = \mathbb{E}\left[\sum_{i=1}^n Z_i^2\right] = \sum_{i=1}^n \mathbb{E}[Z_i^2]$$

Comparing the results, we see that $\mathbb{E}[||\boldsymbol{Z}||^2] = \text{tr}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T])$.

(2) If $\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{0}$, the variance-covariance matrix is defined as:

$$\text{Var}(\boldsymbol{Z}) = \mathbb{E}[(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])^T] = \mathbb{E}[(\boldsymbol{Z} - \boldsymbol{0})(\boldsymbol{Z} - \boldsymbol{0})^T] = \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T]$$

Substituting this into the result from part (1):

$$\mathbb{E}[||\boldsymbol{Z}||^2] = \text{tr}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T]) = \text{tr}(\text{Var}(\boldsymbol{Z}))$$

This useful identity connects the expected squared length of a mean-zero random vector to the sum of its variances and covariances (specifically, the sum of variances, which are the diagonal elements of the trace).

Example 4.2 (Expected Inner Product and Trace of Covariance)**.** Let $\boldsymbol{U}, \boldsymbol{V}$ be random vectors in $\mathbb{R}^n$. Assume that the expectation of at least one of them is the zero vector (e.g., $\mathbb{E}[\boldsymbol{U}] = \boldsymbol{0}$). Show that:

$$\mathbb{E}[\boldsymbol{U}^T\boldsymbol{V}] = \text{tr}(\text{Cov}(\boldsymbol{V}, \boldsymbol{U}))$$

(Note: The source had $\text{tr}(\text{cov}(VU^T))$, which might be a typo. We use the standard definition $\text{Cov}(\boldsymbol{V}, \boldsymbol{U}) = \mathbb{E}[(\boldsymbol{V} - \mathbb{E}\,\boldsymbol{V})(\boldsymbol{U} - \mathbb{E}\,\boldsymbol{U})^T]$).

Solution: Assume, without loss of generality, that $\mathbb{E}[\boldsymbol{U}] = \boldsymbol{0}$. The inner product $\boldsymbol{U}^T\boldsymbol{V}$ is a scalar (1×1 matrix). A scalar is equal to its trace.

$$\boldsymbol{U}^T\boldsymbol{V} = \text{tr}(\boldsymbol{U}^T\boldsymbol{V})$$

However, the trace is usually applied to square matrices. We can use the property $\text{tr}(AB) = \text{tr}(BA)$ for compatible matrices. Let's rewrite the inner product using a trace:

$$\boldsymbol{U}^T\boldsymbol{V} = \sum_{i=1}^{n} U_i V_i$$

Consider the trace of $\boldsymbol{V}\boldsymbol{U}^T$ (which is $n \times n$):

$$\text{tr}(\boldsymbol{V}\boldsymbol{U}^T) = \sum_{i=1}^{n} (\boldsymbol{V}\boldsymbol{U}^T)_{ii} = \sum_{i=1}^{n} V_i U_i = \boldsymbol{U}^T\boldsymbol{V}$$

So, $\boldsymbol{U}^T\boldsymbol{V} = \text{tr}(\boldsymbol{V}\boldsymbol{U}^T)$. Now take the expectation:

$$\mathbb{E}[\boldsymbol{U}^T\boldsymbol{V}] = \mathbb{E}[\text{tr}(\boldsymbol{V}\boldsymbol{U}^T)] = \text{tr}(\mathbb{E}[\boldsymbol{V}\boldsymbol{U}^T]) \quad \text{(Linearity of Trace and Expectation)}$$

Now let's look at the definition of $\text{Cov}(\boldsymbol{V}, \boldsymbol{U})$:

$$\text{Cov}(\boldsymbol{V}, \boldsymbol{U}) = \mathbb{E}[(\boldsymbol{V} - \mathbb{E}[\boldsymbol{V}])(\boldsymbol{U} - \mathbb{E}[\boldsymbol{U}])^T]$$

Since we assumed $\mathbb{E}[\boldsymbol{U}] = \boldsymbol{0}$:

$$\begin{aligned} \text{Cov}(\boldsymbol{V}, \boldsymbol{U}) &= \mathbb{E}[(\boldsymbol{V} - \mathbb{E}[\boldsymbol{V}])(\boldsymbol{U} - \boldsymbol{0})^T] = \mathbb{E}[(\boldsymbol{V} - \mathbb{E}[\boldsymbol{V}])\boldsymbol{U}^T] \\ &= \mathbb{E}[\boldsymbol{V}\boldsymbol{U}^T - \mathbb{E}[\boldsymbol{V}]\boldsymbol{U}^T] = \mathbb{E}[\boldsymbol{V}\boldsymbol{U}^T] - \mathbb{E}[\mathbb{E}[\boldsymbol{V}]\boldsymbol{U}^T] \\ &= \mathbb{E}[\boldsymbol{V}\boldsymbol{U}^T] - \mathbb{E}[\boldsymbol{V}]\,\mathbb{E}[\boldsymbol{U}^T] \quad \text{(Since } \mathbb{E}[\boldsymbol{V}] \text{ is constant)} \end{aligned}$$

Since $\mathbb{E}[\boldsymbol{U}] = \boldsymbol{0}$, its transpose $\mathbb{E}[\boldsymbol{U}^T] = (\mathbb{E}[\boldsymbol{U}])^T = \boldsymbol{0}^T$.

$$\text{Cov}(\boldsymbol{V}, \boldsymbol{U}) = \mathbb{E}[\boldsymbol{V}\boldsymbol{U}^T] - \mathbb{E}[\boldsymbol{V}]\boldsymbol{0}^T = \mathbb{E}[\boldsymbol{V}\boldsymbol{U}^T]$$

Therefore,

$$\text{tr}(\text{Cov}(\boldsymbol{V}, \boldsymbol{U})) = \text{tr}(\mathbb{E}[\boldsymbol{V}\boldsymbol{U}^T])$$

Comparing this with our expression for $\mathbb{E}[\boldsymbol{U}^T\boldsymbol{V}]$, we conclude:

$$\mathbb{E}[\boldsymbol{U}^T\boldsymbol{V}] = \text{tr}(\text{Cov}(\boldsymbol{V}, \boldsymbol{U}))$$

4.1 Problems from Past Exams

The following problems are adapted from previous exams and test understanding of the linear model's properties, particularly relating to projections, residuals, and covariances.

Example 4.3 (Exam Problem 1 - Moed Aleph 5784 / 2023-24)**.** Assume the standard linear model assumptions hold: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$. Let $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ be the vector of fitted values (where $\hat{\boldsymbol{\beta}}$ is the OLS estimator) and $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ be the vector of residuals. Let $\mathcal{C} = \mathrm{Im}(\boldsymbol{X})$ be the column space of $\boldsymbol{X}$. Recall $\hat{\boldsymbol{Y}} = P_{\mathcal{C}}\boldsymbol{Y}$ where $P_{\mathcal{C}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the orthogonal projection matrix onto $\mathcal{C}$. Assume $\boldsymbol{X}$ has full column rank $k = p + 1$.

(a) Let $\boldsymbol{X}'$ be the matrix obtained by deleting some columns from $\boldsymbol{X}$, and let $L = \mathrm{Im}(\boldsymbol{X}')$ be its column space. Note that $L \subseteq \mathcal{C}$. Let P_L be the orthogonal projection matrix onto L. Show that $P_L\hat{\boldsymbol{Y}} = P_L\boldsymbol{Y}$. Does this result require the linear model assumptions or the normal linear model assumptions?

(b) Find $\mathbb{E}[||\boldsymbol{e}||^2]$.

(c) Calculate the following quantities and state their dimensions: $\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e})$, $\mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{e})$, $\mathrm{Cov}(\boldsymbol{e})$.

(d) Find $\mathrm{Var}(e_i)$, the variance of the i-th residual.

Solution:

(a) We are given $\hat{\boldsymbol{Y}} = P_{\mathcal{C}}\boldsymbol{Y}$ and $L \subseteq \mathcal{C}$. We want to show $P_L\hat{\boldsymbol{Y}} = P_L\boldsymbol{Y}$. Substitute the expression for $\hat{\boldsymbol{Y}}$:

$$P_L\hat{\boldsymbol{Y}} = P_L(P_{\mathcal{C}}\boldsymbol{Y}) = (P_L P_{\mathcal{C}})\boldsymbol{Y}$$

A fundamental property of orthogonal projections is that if $L \subseteq \mathcal{C}$, then projecting onto the larger space $\mathcal{C}$ first and then onto the smaller space L is equivalent to just projecting onto L. That is, $P_L P_{\mathcal{C}} = P_L$. Therefore,

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

This result is purely a property of geometry and linear algebra concerning orthogonal projections onto nested subspaces. It does *not* depend on any statistical assumptions of the linear model (linearity, error distribution, etc.).

(b) We need $\mathbb{E}[||\boldsymbol{e}||^2]$. First, express $\boldsymbol{e}$ in terms of $\boldsymbol{\epsilon}$:

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - P_{\mathcal{C}}\boldsymbol{Y} = (\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{Y}$$

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

$$\boldsymbol{e} = (\boldsymbol{I} - P_{\mathcal{C}})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{\epsilon}$$

Since the columns of $\boldsymbol{X}$ are in $\mathcal{C}$, $P_{\mathcal{C}}\boldsymbol{X} = \boldsymbol{X}$. Thus, $(\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{X} = \boldsymbol{X} - P_{\mathcal{C}}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \mathbf{0}$. So, $\boldsymbol{e} = (\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{\epsilon}$. Now consider $||\boldsymbol{e}||^2 = \boldsymbol{e}^T\boldsymbol{e}$:

$$||\boldsymbol{e}||^2 = ((\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{\epsilon})^T((\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{\epsilon}) = \boldsymbol{\epsilon}^T(\boldsymbol{I} - P_{\mathcal{C}})^T(\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{\epsilon}$$

Since $(\boldsymbol{I} - P_{\mathcal{C}})$ is a projection matrix, it is symmetric and idempotent: $(\boldsymbol{I} - P_{\mathcal{C}})^T = (\boldsymbol{I} - P_{\mathcal{C}})$ and $(\boldsymbol{I} - P_{\mathcal{C}})^2 = (\boldsymbol{I} - P_{\mathcal{C}})$.

$$||\boldsymbol{e}||^2 = \boldsymbol{\epsilon}^T(\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{\epsilon}$$

This is a quadratic form in $\boldsymbol{\epsilon}$. We use the formula $\mathbb{E}[\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}] = \mathrm{tr}(\boldsymbol{A}\,\mathrm{Var}(\boldsymbol{z})) + (\mathbb{E}[\boldsymbol{z}])^T\boldsymbol{A}(\mathbb{E}[\boldsymbol{z}])$. Here, $\boldsymbol{z} = \boldsymbol{\epsilon}$, $\boldsymbol{A} = \boldsymbol{I} - P_{\mathcal{C}}$, $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$.

$$\mathbb{E}[||\boldsymbol{e}||^2] = \mathrm{tr}((\boldsymbol{I} - P_{\mathcal{C}})(\sigma^2\boldsymbol{I}_n)) + \mathbf{0}^T(\boldsymbol{I} - P_{\mathcal{C}})\mathbf{0}$$

$$= \sigma^2\,\mathrm{tr}(\boldsymbol{I} - P_{\mathcal{C}}) + 0$$

The trace of a projection matrix is its rank. $\text{rank}(\boldsymbol{I}) = n$. $\text{rank}(P_{\mathcal{C}}) = \text{rank}(\boldsymbol{X}) = k$ (since $\boldsymbol{X}$ has full column rank $k = p + 1$).

$$\text{tr}(\boldsymbol{I} - P_{\mathcal{C}}) = \text{tr}(\boldsymbol{I}) - \text{tr}(P_{\mathcal{C}}) = n - \text{rank}(P_{\mathcal{C}}) = n - k$$

Therefore,

$$\mathbb{E}[||\boldsymbol{e}||^2] = \sigma^2(n - k)$$

This justifies why $s^2 = ||\boldsymbol{e}||^2/(n - k)$ is an unbiased estimator for σ^2.

(c) We calculate the covariances.

- $\text{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e})$: We have $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) = \boldsymbol{\beta}+(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}$. So $\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}$ (since $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$). And $\boldsymbol{e} - \mathbb{E}[\boldsymbol{e}] = \boldsymbol{e} = (\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{\epsilon}$ (since $\mathbb{E}[\boldsymbol{e}] = \boldsymbol{0}$).

$$\begin{aligned}
\text{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) &= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}])(\boldsymbol{e} - \mathbb{E}[\boldsymbol{e}])^T] \\
&= \mathbb{E}[((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon})((\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{\epsilon})^T] \\
&= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T(\boldsymbol{I} - P_{\mathcal{C}})^T] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T](\boldsymbol{I} - P_{\mathcal{C}}) \quad \text{(constants out, } \boldsymbol{I} - P_{\mathcal{C}} \text{ symmetric)} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma^2\boldsymbol{I}_n)(\boldsymbol{I} - P_{\mathcal{C}}) \quad \text{(since } \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}, \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \text{Var}(\boldsymbol{\epsilon})) \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{I} - P_{\mathcal{C}})
\end{aligned}$$

  We know $(\boldsymbol{I} - P_{\mathcal{C}})$ projects onto the space orthogonal to $\mathcal{C} = \text{Im}(\boldsymbol{X})$. The rows of $\boldsymbol{X}^T$ are in the orthogonal complement of this space. Alternatively, recall $(\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{X} = \boldsymbol{0}$, so $\boldsymbol{X}^T(\boldsymbol{I} - P_{\mathcal{C}}) = (\boldsymbol{X}^T(\boldsymbol{I} - P_{\mathcal{C}})^T)^T = (((\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{X})^T)^T = (\boldsymbol{0}^T)^T = \boldsymbol{0}$. Thus, $\text{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{0} = \boldsymbol{0}$. The dimension is $(k \times n)$. The OLS estimator and the residual vector are uncorrelated.

- $\text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{e})$: $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{e}] = \boldsymbol{0}$.

$$\begin{aligned}
\text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{e}) &= \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{e}^T] = \mathbb{E}[\boldsymbol{\epsilon}((\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{\epsilon})^T] \\
&= \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T(\boldsymbol{I} - P_{\mathcal{C}})^T] = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T](\boldsymbol{I} - P_{\mathcal{C}}) \\
&= (\sigma^2\boldsymbol{I}_n)(\boldsymbol{I} - P_{\mathcal{C}}) = \sigma^2(\boldsymbol{I} - P_{\mathcal{C}})
\end{aligned}$$

  The dimension is $(n \times n)$.

- $\text{Cov}(\boldsymbol{e})$: This is just $\text{Var}(\boldsymbol{e})$.

$$\begin{aligned}
\text{Var}(\boldsymbol{e}) &= \text{Var}((\boldsymbol{I} - P_{\mathcal{C}})\boldsymbol{\epsilon}) \\
&= (\boldsymbol{I} - P_{\mathcal{C}})\,\text{Var}(\boldsymbol{\epsilon})(\boldsymbol{I} - P_{\mathcal{C}})^T \\
&= (\boldsymbol{I} - P_{\mathcal{C}})(\sigma^2\boldsymbol{I}_n)(\boldsymbol{I} - P_{\mathcal{C}}) \\
&= \sigma^2(\boldsymbol{I} - P_{\mathcal{C}})(\boldsymbol{I} - P_{\mathcal{C}}) \\
&= \sigma^2(\boldsymbol{I} - P_{\mathcal{C}}) \quad \text{(Idempotence)}
\end{aligned}$$

  The dimension is $(n \times n)$. Notice $\text{Var}(\boldsymbol{e}) = \text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{e})$.

(d) $\text{Var}(e_i)$ is the i-th diagonal element of the matrix $\text{Var}(\boldsymbol{e})$.

$$\text{Var}(e_i) = (\text{Var}(\boldsymbol{e}))_{ii} = (\sigma^2(\boldsymbol{I} - P_{\mathcal{C}}))_{ii} = \sigma^2(\boldsymbol{I}_{ii} - (P_{\mathcal{C}})_{ii})$$

$$= \sigma^2(1 - (P_{\mathcal{C}})_{ii})$$

The diagonal elements of the projection matrix $P_{\mathcal{C}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ are often called the *leverage values*, denoted h_{ii}. So,

$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$$

The variance of a residual depends on the corresponding leverage value h_{ii}, which measures how influential the i-th observation is in determining the fit.

Example 4.4 (Exam Problem 2 - Midterm Exam 5762 / 2001-02, Generalized Least Squares)**.** Consider a linear model $\boldsymbol{Y} = \boldsymbol{X\beta}+\boldsymbol{\epsilon}$, where the errors satisfy the standard assumptions $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$. Let $\boldsymbol{\Sigma}$ be a known $n \times n$ symmetric positive definite matrix.

For any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, define the *Mahalanobis norm* induced by $\boldsymbol{\Sigma}^{-1}$ as:

$$||\boldsymbol{u} - \boldsymbol{v}||^2_{\Sigma^{-1}} = (\boldsymbol{u} - \boldsymbol{v})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{u} - \boldsymbol{v})$$

Define the *Generalized Least Squares (GLS)* estimator $\hat{\boldsymbol{\beta}}_\Sigma$ as the vector minimizing this norm between $\boldsymbol{Y}$ and $\boldsymbol{Xb}$:

$$\hat{\boldsymbol{\beta}}_\Sigma = \mathrm{argmin}_{\boldsymbol{b}\in\mathbb{R}^k}\, ||\boldsymbol{Y} - \boldsymbol{Xb}||^2_{\Sigma^{-1}}$$

(a) Show that

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

Explain what estimator is obtained in the special case where $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_n$. (Hint: Since $\boldsymbol{\Sigma}$ is symmetric positive definite, so is $\boldsymbol{\Sigma}^{-1}$. Thus, there exists a matrix $\boldsymbol{C}$ such that $\boldsymbol{C}^T\boldsymbol{C} = \boldsymbol{\Sigma}^{-1}$. Consider transforming the variables using $\boldsymbol{C}$.)

(b) Assume now that the true model is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ with $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$ but $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$ (where $\boldsymbol{\Sigma}$ is known and positive definite, possibly different from $\sigma^2\boldsymbol{I}$). Show that the GLS estimator $\hat{\boldsymbol{\beta}}_\Sigma$ found in part (a) is unbiased for $\boldsymbol{\beta}$, and find its variance-covariance matrix, $\mathrm{Var}(\hat{\boldsymbol{\beta}}_\Sigma)$.

Solution:

(a) We want to minimize $S(\boldsymbol{b}) = (\boldsymbol{Y} - \boldsymbol{Xb})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{Xb})$. Following the hint, since $\boldsymbol{\Sigma}^{-1}$ is symmetric positive definite, we can perform a Cholesky decomposition $\boldsymbol{\Sigma}^{-1} = \boldsymbol{LL}^T$ where $\boldsymbol{L}$ is lower triangular and invertible, or more generally find a matrix $\boldsymbol{C}$ (e.g., the symmetric positive definite square root $\boldsymbol{\Sigma}^{-1/2}$) such that $\boldsymbol{C}^T\boldsymbol{C} = \boldsymbol{\Sigma}^{-1}$. Let's use such a $\boldsymbol{C}$. We can rewrite the objective function:

$$\begin{aligned}
S(\boldsymbol{b}) &= (\boldsymbol{Y} - \boldsymbol{Xb})^T\boldsymbol{C}^T\boldsymbol{C}(\boldsymbol{Y} - \boldsymbol{Xb}) \\
&= (\boldsymbol{C}(\boldsymbol{Y} - \boldsymbol{Xb}))^T(\boldsymbol{C}(\boldsymbol{Y} - \boldsymbol{Xb})) \\
&= (\boldsymbol{CY} - \boldsymbol{CXb})^T(\boldsymbol{CY} - \boldsymbol{CXb})
\end{aligned}$$

Let's define transformed variables: $\tilde{\boldsymbol{Y}} = \boldsymbol{CY}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{CX}$. Then the objective function becomes:

$$S(\boldsymbol{b}) = (\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{b})^T(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{b}) = ||\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{b}||^2$$

This is now a standard OLS problem for the transformed model $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\boldsymbol{b} + \tilde{\boldsymbol{\epsilon}}$ (where $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{C\epsilon}$). The value of $\boldsymbol{b}$ that minimizes this sum of squares is the OLS solution for the transformed system:

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

Now substitute back the original variables:

- $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} = (\boldsymbol{CX})^T(\boldsymbol{CX}) = \boldsymbol{X}^T\boldsymbol{C}^T\boldsymbol{CX} = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$
- $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}} = (\boldsymbol{CX})^T(\boldsymbol{CY}) = \boldsymbol{X}^T\boldsymbol{C}^T\boldsymbol{CY} = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$

So, the minimizing vector $\hat{\boldsymbol{\beta}}_\Sigma$ is:

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

Special Case: If $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_n$, then $\boldsymbol{\Sigma}^{-1} = (\sigma^2\boldsymbol{I}_n)^{-1} = \frac{1}{\sigma^2}\boldsymbol{I}_n$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\sigma^2\boldsymbol{I}} &= (\boldsymbol{X}^T(\frac{1}{\sigma^2}\boldsymbol{I})\boldsymbol{X})^{-1}\boldsymbol{X}^T(\frac{1}{\sigma^2}\boldsymbol{I})\boldsymbol{Y} \\
&= (\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X})^{-1}(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y}) \\
&= (\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y}) \quad \text{(using } (cA)^{-1} = c^{-1}A^{-1} \text{ for scalar } c\text{)} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}
\end{aligned}$$

This is exactly the formula for the Ordinary Least Squares (OLS) estimator $\hat{\boldsymbol{\beta}}_{OLS}$. Minimizing the Mahalanobis norm with $\boldsymbol{\Sigma}^{-1}$ proportional to the identity matrix is equivalent to OLS.

(b) Now we assume the true model has $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$ and $\text{Var}(\boldsymbol{Y}) = \text{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$. We want to find the expectation and variance of $\hat{\boldsymbol{\beta}}_\Sigma = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$. Let $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}$. This is a constant matrix (treating $\boldsymbol{X}$ and $\boldsymbol{\Sigma}$ as fixed). Then $\hat{\boldsymbol{\beta}}_\Sigma = \boldsymbol{A}\boldsymbol{Y}$.

Unbiasedness:

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}_\Sigma] &= \mathbb{E}[\boldsymbol{A}\boldsymbol{Y}] = \boldsymbol{A}\,\mathbb{E}[\boldsymbol{Y}] \\
&= (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})\boldsymbol{\beta} \\
&= \boldsymbol{I}\boldsymbol{\beta} = \boldsymbol{\beta}
\end{aligned}$$

So, $\hat{\boldsymbol{\beta}}_\Sigma$ is an unbiased estimator for $\boldsymbol{\beta}$ under the assumption that $\text{Var}(\boldsymbol{Y}) = \boldsymbol{\Sigma}$.

Variance-Covariance Matrix:

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}_\Sigma) &= \text{Var}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\,\text{Var}(\boldsymbol{Y})\boldsymbol{A}^T \\
&= \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T \\
&= [(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}]\boldsymbol{\Sigma}[(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}]^T
\end{aligned}$$

Let's find $\boldsymbol{A}^T$:

$$\begin{aligned}
\boldsymbol{A}^T &= ((\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1})^T \\
&= (\boldsymbol{\Sigma}^{-1})^T(\boldsymbol{X}^T)^T((\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1})^T \\
&= \boldsymbol{\Sigma}^{-1}\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^T)^{-1} \quad \text{(since } \boldsymbol{\Sigma}^{-1} \text{ is symmetric)} \\
&= \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T(\boldsymbol{\Sigma}^{-1})^T(\boldsymbol{X}^T)^T)^{-1} \\
&= \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \quad \text{(since } \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \text{ is symmetric)}
\end{aligned}$$

Now substitute $\boldsymbol{A}$ and $\boldsymbol{A}^T$ into the variance formula:

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}_\Sigma) &= [(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}]\boldsymbol{\Sigma}[\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}] \\
&= (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{I}\boldsymbol{\Sigma}^{-1})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \\
&= \boldsymbol{I}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}
\end{aligned}$$

Thus, the variance-covariance matrix of the GLS estimator is $\mathrm{Var}(\hat{\boldsymbol{\beta}}_\Sigma) = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$. This famous result shows that using the inverse of the true covariance matrix as the weight matrix in the least squares criterion leads to a simple expression for the variance of the resulting estimator. The Gauss-Markov theorem further states that this GLS estimator is the Best Linear Unbiased Estimator (BLUE) when $\mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$.

Administrative Note: Please remember to review the properties of covariance matrices (listed in Section 1) as they will be part of your upcoming homework assignment. Ensure you understand the derivations from the examples, particularly those involving projections and the properties of OLS/GLS estimators. The distinction between model assumptions and derived results (Example 2.2) is crucial.