Lecture Notes: Covariance Matrices and the Linear Model

Undergraduate Mathematics Educator

Based on Lecture Transcript

Administrative Notes and Announcements

- **Recitation Material:** Please note that the properties of covariance matrices discussed at the beginning of this lecture were covered thoroughly, including proofs. I have asked Niv (the TA) *not* to repeat this specific material in the recitation sections to allow more time for problem-solving. If you have further questions on these properties after reviewing your notes, please utilize office hours (either mine or Niv's).
- **Office Hours:** Please make use of office hours if you have questions about the material or homework. Thinking through the concepts at home first and then bringing specific questions is often very productive.
- **Context from Previous Lecture:** We are picking up directly from where we left off before the Passover break, returning to the linear statistical model.
- **Assumption on Regressors:** A point clarified during the lecture: Throughout this course, unless explicitly stated otherwise, we will treat the predictor variables (the entries in the matrix $\mathbf{X}$) as **fixed, non-random constants**. The randomness in our model comes entirely from the error term $\boldsymbol{\epsilon}$. This is a standard assumption in many regression contexts. It means our X values are considered given data points, not outcomes of a random process.

---

1 Properties of Covariance Matrices: Revisited

We began by revisiting some key properties of covariance matrices for random vectors, picking up exactly where we left off last time. We had already established the first few properties. Let's focus on Properties 3 through 6, paying special attention to an alternative proof for Property 3.

Let $\boldsymbol{Z}$ and $\boldsymbol{W}$ be random vectors, and let $\mathbf{A}$ and $\mathbf{B}$ be constant (non-random) matrices of appropriate dimensions.

Property 1.1 (Covariance of Linear Transformations - Property 3 Revisited)**.**

$$Cov(\mathbf{A}\boldsymbol{Z}, \mathbf{B}\boldsymbol{W}) = \mathbf{A}\,Cov(\boldsymbol{Z}, \boldsymbol{W})\mathbf{B}^T$$

Alternative Proof using Matrix Identity. Recall the definition of the covariance matrix:

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

Let $\boldsymbol{U} = \mathbf{A}\boldsymbol{Z}$ and $\boldsymbol{V} = \mathbf{B}\boldsymbol{W}$. We need to compute the two terms on the right-hand side for these specific $\boldsymbol{U}$ and $\boldsymbol{V}$.

First Term: $\mathbb{E}[\boldsymbol{U}\boldsymbol{V}^T]$

$$\begin{aligned}\mathbb{E}[(\mathbf{A}\boldsymbol{Z})(\mathbf{B}\boldsymbol{W})^T] &= \mathbb{E}[\mathbf{A}\boldsymbol{Z}\boldsymbol{W}^T\mathbf{B}^T] \\ &= \mathbf{A}\mathbb{E}[\boldsymbol{Z}\boldsymbol{W}^T]\mathbf{B}^T\end{aligned}$$

Here, we used the linearity of expectation and the property that constant matrices can be factored out. Remember, $\mathbf{A}$ and $\mathbf{B}$ are constant matrices, while $\boldsymbol{Z}$ and $\boldsymbol{W}$ are random vectors, making $\boldsymbol{Z}\boldsymbol{W}^T$ a random matrix.

Second Term: $\mathbb{E}[\boldsymbol{U}]\mathbb{E}[\boldsymbol{V}]^T$

$$\begin{aligned}\mathbb{E}[\mathbf{A}\boldsymbol{Z}](\mathbb{E}[\mathbf{B}\boldsymbol{W}])^T &= (\mathbf{A}\mathbb{E}[\boldsymbol{Z}])(\mathbf{B}\mathbb{E}[\boldsymbol{W}])^T \\ &= (\mathbf{A}\mathbb{E}[\boldsymbol{Z}])(\mathbb{E}[\boldsymbol{W}]^T\mathbf{B}^T) \\ &= \mathbf{A}\mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{W}]^T\mathbf{B}^T\end{aligned}$$

Again, we used the linearity of expectation. Let $\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{\mu}_Z$ and $\mathbb{E}[\boldsymbol{W}] = \boldsymbol{\mu}_W$.

Combining Terms:

$$\begin{aligned}\mathrm{Cov}(\mathbf{A}\boldsymbol{Z}, \mathbf{B}\boldsymbol{W}) &= \mathbb{E}[(\mathbf{A}\boldsymbol{Z})(\mathbf{B}\boldsymbol{W})^T] - \mathbb{E}[\mathbf{A}\boldsymbol{Z}]\mathbb{E}[\mathbf{B}\boldsymbol{W}]^T \\ &= (\mathbf{A}\mathbb{E}[\boldsymbol{Z}\boldsymbol{W}^T]\mathbf{B}^T) - (\mathbf{A}\boldsymbol{\mu}_Z\boldsymbol{\mu}_W^T\mathbf{B}^T) \\ &= \mathbf{A}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{W}^T] - \boldsymbol{\mu}_Z\boldsymbol{\mu}_W^T)\mathbf{B}^T \\ &= \mathbf{A}\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{W})\mathbf{B}^T\end{aligned}$$

This completes the alternative proof using the matrix identity. It's a valuable exercise to work through this algebra to become comfortable with manipulating expectations and transposes involving matrices and vectors. □

Remark 1.2. Understanding both the entry-wise proof (from the previous lecture) and this matrix-based proof deepens our understanding of covariance.

Property 1.3 (Variance of a Linear Transformation - Property 4)**.** *If $Var(\boldsymbol{Z})$ denotes the covariance matrix $Cov(\boldsymbol{Z}, \boldsymbol{Z})$, then*

$$Var(\mathbf{A}\boldsymbol{Z}) = \mathbf{A}\, Var(\boldsymbol{Z})\mathbf{A}^T$$

Proof. This is a direct special case of Property 3 where we set $\mathbf{B} = \mathbf{A}$ and $\boldsymbol{W} = \boldsymbol{Z}$.

$$\mathrm{Cov}(\mathbf{A}\boldsymbol{Z}, \mathbf{A}\boldsymbol{Z}) = \mathbf{A}\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{Z})\mathbf{A}^T = \mathbf{A}\mathrm{Var}(\boldsymbol{Z})\mathbf{A}^T$$

□

Remark 1.4 (The "Butterfly Formula")**.** Informally, Property 4 is sometimes called the "butterfly formula" because the matrix $\mathbf{A}$ appears on the left and its transpose $\mathbf{A}^T$ appears on the right, like wings around the original variance matrix $\mathrm{Var}(\boldsymbol{Z})$. This transformation rule is extremely useful.

Property 1.5 (Variance of an Inner Product - Property 5)**.** *Let $\boldsymbol{a}$ be a constant vector. Then $\boldsymbol{a}^T\boldsymbol{Z}$ is a scalar random variable, and its variance is given by:*

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

Proof. First, note that $\boldsymbol{a}^T\boldsymbol{Z}$ is indeed a scalar random variable, as it's the inner product of a constant vector $\boldsymbol{a}$ and a random vector $\boldsymbol{Z}$. We can view the row vector $\boldsymbol{a}^T$ as a $1 \times n$ matrix (if $\boldsymbol{Z}$ is n-dimensional). Applying Property 4 with $\mathbf{A} = \boldsymbol{a}^T$:

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

Note that the result $\boldsymbol{a}^T\text{Var}(\boldsymbol{Z})\boldsymbol{a}$ is a 1×1 matrix, which is equivalent to a scalar, as expected for a variance. □

Remark 1.6 (Quadratic Forms)**.** The expression $\boldsymbol{a}^T\text{Var}(\boldsymbol{Z})\boldsymbol{a}$ is an example of a **quadratic form**. If $\boldsymbol{\Sigma} = \text{Var}(\boldsymbol{Z})$ is an $n \times n$ matrix, then for a vector $\boldsymbol{x} \in \mathbb{R}^n$, the function $f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x}$ is called a quadratic form. Explicitly,

$$\boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n}\Sigma_{ij}x_ix_j$$

Ensure you are comfortable with this expansion; it involves basic matrix multiplication.

Property 1.7 (Positive Semidefiniteness - Property 6)**.** *The covariance matrix Var($\boldsymbol{Z}$) is always positive semidefinite.*

Proof. A matrix $\boldsymbol{\Sigma}$ is positive semidefinite if it is symmetric and $\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a} \geq 0$ for all conformable vectors $\boldsymbol{a}$.

First, we establish symmetry. $\text{Var}(\boldsymbol{Z}) = \text{Cov}(\boldsymbol{Z}, \boldsymbol{Z})$. From Property 1 (which stated $\text{Cov}(\boldsymbol{U}, \boldsymbol{V}) = \text{Cov}(\boldsymbol{V}, \boldsymbol{U})^T$), we have $\text{Var}(\boldsymbol{Z}) = \text{Cov}(\boldsymbol{Z}, \boldsymbol{Z}) = \text{Cov}(\boldsymbol{Z}, \boldsymbol{Z})^T = \text{Var}(\boldsymbol{Z})^T$. So, $\text{Var}(\boldsymbol{Z})$ is symmetric. (Note: Covariance $\text{Cov}(\boldsymbol{Z}, \boldsymbol{W})$ is generally not square or symmetric unless $\boldsymbol{Z}$ and $\boldsymbol{W}$ have the same dimension).

Second, we need to show $\boldsymbol{a}^T\text{Var}(\boldsymbol{Z})\boldsymbol{a} \geq 0$ for any constant vector $\boldsymbol{a}$. From Property 5, we know:

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

Since the variance of any scalar random variable (like $\boldsymbol{a}^T\boldsymbol{Z}$) must be non-negative, we have:

$$\text{Var}(\boldsymbol{a}^T\boldsymbol{Z}) \geq 0$$

Therefore, $\boldsymbol{a}^T\text{Var}(\boldsymbol{Z})\boldsymbol{a} \geq 0$ for all $\boldsymbol{a}$, confirming that $\text{Var}(\boldsymbol{Z})$ is positive semidefinite. □

2 The Linear Statistical Model

Having refreshed these essential properties of covariance matrices, let's return to the primary object of study: the linear statistical model. We introduced this before the break, but let's restate it carefully.

Definition 2.1 (Linear Model - Scalar Form)**.** For n observations, indexed by $i = 1, \ldots, n$, the model is given by:

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

Or, more compactly using summation notation and letting $x_{i0} = 1$:

$$Y_i = \sum_{j=0}^{p}\beta_j x_{ij} + \epsilon_i \quad \text{for } i = 1, \ldots, n$$

Where:

- Y_i is the response variable for the i-th observation.
- x_{ij} is the value of the j-th predictor variable for the i-th observation (with $x_{i0} = 1$ for the intercept).
- β_j are the unknown model parameters (coefficients).
- ϵ_i is the random error term for the i-th observation.

The key assumptions on the error terms ϵ_i are:

1. **Zero Mean:** $\mathbb{E}[\epsilon_i] = 0$ for all i.
2. **Constant Variance (Homoscedasticity):** $\text{Var}(\epsilon_i) = \sigma^2$ for all i, where $\sigma^2 > 0$ is an unknown parameter.
3. **Uncorrelated Errors:** $\text{Cov}(\epsilon_i, \epsilon_k) = 0$ for all $i \neq k$.

Assumption 2.2 (Fixed Regressors)**.** As mentioned in the administrative notes, we assume the predictor values x_{ij} (for $j = 1, \ldots, p$ and $i = 1, \ldots, n$) are **fixed, known constants**. They are not random variables. The only source of randomness in Y_i comes from the error term ϵ_i.

Remark 2.3. Under the fixed regressors assumption, the conditioning on X values we discussed previously becomes implicit. The properties of the errors (mean 0, variance σ^2, uncorrelated) hold unconditionally in this framework.

Definition 2.4 (Linear Model - Matrix Form)**.** The scalar equations for all n observations can be written elegantly in matrix form:

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

Where:

- $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ is the $n \times 1$ vector of responses.
- $\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}$ is the $n \times (p+1)$ **design matrix** (or model matrix), assumed to be known and constant.
- $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ is the $(p+1) \times 1$ vector of unknown parameters.
- $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ is the $n \times 1$ vector of random errors.

The assumptions on the error vector $\boldsymbol{\epsilon}$ translate to:

1. $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ (where $\mathbf{0}$ is the $n \times 1$ zero vector).
2. $\text{Var}(\boldsymbol{\epsilon}) = \text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$ (where $\mathbf{I}_n$ is the $n \times n$ identity matrix). This matrix has σ^2 on the diagonal (capturing $\text{Var}(\epsilon_i) = \sigma^2$) and 0s off-diagonal (capturing $\text{Cov}(\epsilon_i, \epsilon_k) = 0$ for $i \neq k$).

The unknown parameters of the model are the vector $\boldsymbol{\beta}$ and the scalar variance σ^2.

Remark 2.5 (Model Scope)**.** This model is quite general. We haven't specified the *distribution* of the errors (e.g., Normal distribution), only their first and second moments (mean and covariance). Many different error distributions could satisfy these conditions. Our goal is typically to perform statistical inference on the parameter vector $\boldsymbol{\beta}$, which describes the relationship between the predictors and the response. σ^2 is often considered a nuisance parameter, though estimating it is also important.

3 Least Squares Estimation Revisited

We previously defined the Least Squares Estimator (LSE) purely from an algebraic/geometric perspective: finding the coefficient vector $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squared differences between observed Y_i and fitted values $\hat{Y}_i$. Now, we interpret this within the context of the linear statistical model.

Definition 3.1 (LSE Components under the Model)**.**

- **LSE Estimator:** $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$. This is now interpreted as an *estimator* for the true, unknown parameter vector $\boldsymbol{\beta}$.
- **Fitted Values:** $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$. Let $\mathbf{P}_M = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. This is the **projection matrix** onto the column space of $\mathbf{X}$ (denoted $M = \text{Im}(\mathbf{X})$). So, $\hat{\boldsymbol{Y}} = \mathbf{P}_M\boldsymbol{Y}$. The vector of fitted values is the orthogonal projection of the observed response vector $\boldsymbol{Y}$ onto the subspace spanned by the columns of the design matrix $\mathbf{X}$. $\hat{\boldsymbol{Y}}$ serves as an estimate for the systematic part $\mathbf{X}\boldsymbol{\beta}$.
- **Residuals:** $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \mathbf{P}_M\boldsymbol{Y} = (\mathbf{I} - \mathbf{P}_M)\boldsymbol{Y}$. Let $\mathbf{Q} = \mathbf{I} - \mathbf{P}_M$. This is the projection matrix onto the orthogonal complement of the column space of $\mathbf{X}$, denoted $M^\perp$. Thus, $\boldsymbol{e} = \mathbf{Q}\boldsymbol{Y}$. The residual vector is the orthogonal projection of $\boldsymbol{Y}$ onto the space orthogonal to the column space of $\mathbf{X}$.

Remark 3.2 (Residuals vs. Errors)**.** It is crucial to distinguish between the **residuals** $\boldsymbol{e}$ and the true **errors** $\boldsymbol{\epsilon}$.

- $\boldsymbol{\epsilon} = \boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}$ represents the deviation of the observations from the true underlying relationship. We typically do not observe $\boldsymbol{\epsilon}$ because $\boldsymbol{\beta}$ is unknown.
- $\boldsymbol{e} = \boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ represents the deviation of the observations from the fitted relationship. We *can* calculate $\boldsymbol{e}$ from the data.

Geometrically, $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The LSE process finds $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ as the projection of $\boldsymbol{Y}$ onto the column space M. The residual vector $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ lies in the orthogonal space $M^\perp$. While $\boldsymbol{e}$ is not $\boldsymbol{\epsilon}$, it serves as our observable proxy or estimate for the unobservable noise. We will see that $\boldsymbol{e} = \mathbf{Q}\boldsymbol{Y} = \mathbf{Q}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{Q}\mathbf{X}\boldsymbol{\beta} + \mathbf{Q}\boldsymbol{\epsilon}$. Since the columns of $\mathbf{X}$ are in M, and $\mathbf{Q}$ projects onto $M^\perp$, $\mathbf{Q}\mathbf{X} = \mathbf{0}$. Therefore, $\boldsymbol{e} = \mathbf{Q}\boldsymbol{\epsilon}$. The residuals are the projection of the true errors onto the space orthogonal to the column space of $\mathbf{X}$.

4 Properties of LSE Estimators under the Linear Model

Now we use the assumptions of the linear model ($\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$) to derive statistical properties of our estimators $\hat{\boldsymbol{\beta}}$ and related quantities.

Proposition 4.1 (Expectation of Y)**.** *Under the linear model assumptions, the expected value of the response vector is:*

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

Proof.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Y}] &= \mathbb{E}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] \\
&= \mathbb{E}[\mathbf{X}\boldsymbol{\beta}] + \mathbb{E}[\boldsymbol{\epsilon}] \quad \text{(Linearity of Expectation)} \\
&= \mathbf{X}\boldsymbol{\beta} + \mathbf{0} \quad \text{(Since } \mathbf{X}, \boldsymbol{\beta} \text{ are constant and } \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}\text{)} \\
&= \mathbf{X}\boldsymbol{\beta}
\end{aligned}$$

The expected value of the observations lies entirely in the systematic part of the model. □

Theorem 4.2 (Unbiasedness of LSE)**.** *Under the linear model assumptions, the LSE $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. That is,*

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

Proof. Let $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Since $\mathbf{X}$ is constant, $\mathbf{A}$ is also a constant matrix.

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}] \\
&= \mathbb{E}[\mathbf{A}\boldsymbol{Y}] \\
&= \mathbf{A}\mathbb{E}[\boldsymbol{Y}] \quad \text{(Linearity of Expectation, } \mathbf{A} \text{ is constant)} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}) \quad \text{(Since } \mathbb{E}[\boldsymbol{Y}] = \mathbf{X}\boldsymbol{\beta}\text{)} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} \\
&= \mathbf{I}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}$$

On average, the LSE procedure correctly identifies the true parameter vector. □

Proposition 4.3 (Covariance Matrix of Y)**.** *Under the linear model assumptions, the covariance matrix of the response vector is:*

$$Var(\boldsymbol{Y}) = \sigma^2\mathbf{I}_n$$

Proof.

$$\begin{aligned}
\text{Var}(\boldsymbol{Y}) &= \text{Var}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= \text{Var}(\boldsymbol{\epsilon}) \quad \text{(Since } \mathbf{X}\boldsymbol{\beta} \text{ is a constant vector)} \\
&= \sigma^2\mathbf{I}_n \quad \text{(Model assumption)}
\end{aligned}$$

The variability and correlation structure of the observations are determined solely by the error term. □

Theorem 4.4 (Covariance Matrix of LSE)**.** *Under the linear model assumptions, the covariance matrix of the LSE $\hat{\boldsymbol{\beta}}$ is:*

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

Proof. Again, let $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. We want to compute $\text{Var}(\hat{\boldsymbol{\beta}}) = \text{Var}(\mathbf{A}\boldsymbol{Y})$. We use the "butterfly formula" (Property 4): $\text{Var}(\mathbf{A}\boldsymbol{Y}) = \mathbf{A}\text{Var}(\boldsymbol{Y})\mathbf{A}^T$.

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}) &= \mathbf{A}\text{Var}(\boldsymbol{Y})\mathbf{A}^T \\
&= [(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T](\sigma^2\mathbf{I}_n)[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\
&= \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{I}_n[\mathbf{X}((\mathbf{X}^T\mathbf{X})^{-1})^T] \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \quad \text{(Since } (\mathbf{X}^T\mathbf{X})^{-1} \text{ is symmetric)} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2\mathbf{I}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}$$

This $(p+1) \times (p+1)$ matrix describes the variances of the individual coefficient estimators $(\hat{\beta}_j)$ on its diagonal and the covariances between different pairs $(\hat{\beta}_j, \hat{\beta}_k)$ off-diagonal. Its magnitude depends on the error variance σ^2 and the structure of the design matrix $\mathbf{X}$ through $(\mathbf{X}^T\mathbf{X})^{-1}$. □

5 Estimating the Error Variance σ^2

While our primary interest is often $\boldsymbol{\beta}$, the error variance σ^2 is also an important unknown parameter. It quantifies the residual variability not explained by the predictors. Furthermore, the precision of $\hat{\boldsymbol{\beta}}$ (as seen in its covariance matrix) depends on σ^2. Thus, we need an estimator for σ^2.

Intuition: We don't observe the true errors ϵ_i. However, we do observe the residuals $e_i = Y_i - \hat{Y}_i$. Since the residuals e_i are our empirical stand-ins for the errors ϵ_i, it seems reasonable to base an estimator for the variance of the ϵ_i's (which is σ^2) on the variability of the residuals e_i. A natural measure of the overall size of the residuals is the sum of squared residuals (SSR), $\|\boldsymbol{e}\|^2 = \sum_{i=1}^n e_i^2$.

Definition 5.1 (Estimator for σ^2)**.** The commonly used unbiased estimator for the error variance σ^2 is:

$$\hat{\sigma}^2 = S^2 = \frac{\|\boldsymbol{e}\|^2}{n-p-1} = \frac{\sum_{i=1}^n e_i^2}{n-p-1} = \frac{\text{SSR}}{n-p-1}$$

The quantity $n-p-1$ is known as the **degrees of freedom for error**.

Theorem 5.2 (Unbiasedness of $\hat{\sigma}^2$)**.** *Under the linear model assumptions, $\hat{\sigma}^2$ is an unbiased estimator of σ^2.*

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2$$

We will present two proofs for this important result.

Proof 1: Direct Calculation using Trace. Our goal is to compute $\mathbb{E}[\hat{\sigma}^2]$. Since $\hat{\sigma}^2 = \frac{1}{n-p-1}\|\boldsymbol{e}\|^2$, and the denominator is constant, this is equivalent to showing $\mathbb{E}[\|\boldsymbol{e}\|^2] = \sigma^2(n-p-1)$.

Recall $\boldsymbol{e} = \mathbf{Q}\boldsymbol{Y} = \mathbf{Q}\boldsymbol{\epsilon}$, where $\mathbf{Q} = \mathbf{I} - \mathbf{P}_M$. The squared norm is $\|\boldsymbol{e}\|^2 = \boldsymbol{e}^T\boldsymbol{e}$.

$$\|\boldsymbol{e}\|^2 = (\mathbf{Q}\boldsymbol{\epsilon})^T(\mathbf{Q}\boldsymbol{\epsilon}) = \boldsymbol{\epsilon}^T\mathbf{Q}^T\mathbf{Q}\boldsymbol{\epsilon}$$

Since $\mathbf{Q}$ is a projection matrix, it is symmetric ($\mathbf{Q}^T = \mathbf{Q}$) and idempotent ($\mathbf{Q}^2 = \mathbf{Q}$). Thus, $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q} = \mathbf{Q}$.

$$\|\boldsymbol{e}\|^2 = \boldsymbol{\epsilon}^T\mathbf{Q}\boldsymbol{\epsilon}$$

This is a quadratic form in the random vector $\boldsymbol{\epsilon}$. Expanding this:

$$\|\boldsymbol{e}\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}\epsilon_i\epsilon_j$$

Now, let's take the expectation:

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{e}\|^2] &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}\epsilon_i\epsilon_j\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}\mathbb{E}[\epsilon_i\epsilon_j] \quad \text{(Linearity of Expectation)}\end{aligned}$$

Recall that $\mathbb{E}[\epsilon_i\epsilon_j] = \text{Cov}(\epsilon_i, \epsilon_j) + \mathbb{E}[\epsilon_i]\mathbb{E}[\epsilon_j]$. Since $\mathbb{E}[\epsilon_i] = 0$ for all i, we have $\mathbb{E}[\epsilon_i\epsilon_j] = \text{Cov}(\epsilon_i, \epsilon_j)$. From the model assumptions, $\text{Cov}(\epsilon_i, \epsilon_j) = \sigma^2$ if $i = j$, and 0 if $i \neq j$. So, the only non-zero terms in the sum occur when $i = j$:

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{e}\|^2] &= \sum_{i=1}^{n} q_{ii}\mathbb{E}[\epsilon_i^2] + \sum_{i\neq j} q_{ij}\mathbb{E}[\epsilon_i\epsilon_j] \\ &= \sum_{i=1}^{n} q_{ii}\text{Var}(\epsilon_i) + \sum_{i\neq j} q_{ij}\cdot 0 \\ &= \sum_{i=1}^{n} q_{ii}\sigma^2 \\ &= \sigma^2\sum_{i=1}^{n} q_{ii}\end{aligned}$$

The sum $\sum_{i=1}^{n} q_{ii}$ is the trace of the matrix $\mathbf{Q}$, denoted $\text{Trace}(\mathbf{Q})$. So,

$$\mathbb{E}[\|\boldsymbol{e}\|^2] = \sigma^2\text{Trace}(\mathbf{Q})$$

What is the trace of $\mathbf{Q} = \mathbf{I}_n - \mathbf{P}_M$? We know $\text{Trace}(\mathbf{A}-\mathbf{B}) = \text{Trace}(\mathbf{A}) - \text{Trace}(\mathbf{B})$. $\text{Trace}(\mathbf{I}_n) = n$. The trace of a projection matrix equals the dimension of the space it projects onto. $\mathbf{P}_M$ projects onto the column space of $\mathbf{X}$, which has dimension equal to the rank of $\mathbf{X}$. Assuming $\mathbf{X}$ has full column rank (which is required for $(\mathbf{X}^T\mathbf{X})^{-1}$ to exist), the rank is $p+1$. So, $\text{Trace}(\mathbf{P}_M) = p+1$. Therefore, $\text{Trace}(\mathbf{Q}) = \text{Trace}(\mathbf{I}_n) - \text{Trace}(\mathbf{P}_M) = n-(p+1) = n-p-1$. Substituting this back:

$$\mathbb{E}[\|\boldsymbol{e}\|^2] = \sigma^2(n-p-1)$$

Finally,

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\left[\frac{\|\boldsymbol{e}\|^2}{n-p-1}\right] = \frac{1}{n-p-1}\mathbb{E}[\|\boldsymbol{e}\|^2] = \frac{1}{n-p-1}\sigma^2(n-p-1) = \sigma^2$$

This confirms that $\hat{\sigma}^2$ is an unbiased estimator for σ^2. □

Before the second proof, we establish a useful general lemma.

Lemma 5.3 (Expectation of Squared Norm and Trace)**.** *Let $\boldsymbol{Z}$ be a random vector with finite dimension, mean $\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{\mu}_Z$, and covariance matrix $Var(\boldsymbol{Z}) = \boldsymbol{\Sigma}_Z$. Then,*

$$\mathbb{E}[\|\boldsymbol{Z}\|^2] = \mathbb{E}[\boldsymbol{Z}^T\boldsymbol{Z}] = Trace(Var(\boldsymbol{Z})) + \|\mathbb{E}[\boldsymbol{Z}]\|^2 = Trace(\boldsymbol{\Sigma}_Z) + \boldsymbol{\mu}_Z^T\boldsymbol{\mu}_Z$$

Alternatively, using $Var(\boldsymbol{Z}) = \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T] - \mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{Z}]^T$, we can write:

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

Proof. Let's prove $\mathbb{E}[\boldsymbol{Z}^T\boldsymbol{Z}] = \text{Trace}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T])$. Note that $\boldsymbol{Z}^T\boldsymbol{Z}$ is a scalar (1×1 matrix). For any scalar s, $s = \text{Trace}(s)$.

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

Using the cyclic property of the trace, $\text{Trace}(AB) = \text{Trace}(BA)$. Let $A = \boldsymbol{Z}^T$ and $B = \boldsymbol{Z}$.

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

Since Trace is a linear operator (sum of diagonal elements), we can swap Trace and Expectation:

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

This proves the second form. For the first form:

$$\begin{aligned}
\text{Trace}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T]) &= \text{Trace}(\text{Var}(\boldsymbol{Z}) + \mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{Z}]^T) \\
&= \text{Trace}(\text{Var}(\boldsymbol{Z})) + \text{Trace}(\boldsymbol{\mu}_Z\boldsymbol{\mu}_Z^T) \\
&= \text{Trace}(\text{Var}(\boldsymbol{Z})) + \text{Trace}(\boldsymbol{\mu}_Z^T\boldsymbol{\mu}_Z) \quad \text{(Cyclic property)} \\
&= \text{Trace}(\text{Var}(\boldsymbol{Z})) + \boldsymbol{\mu}_Z^T\boldsymbol{\mu}_Z \quad \text{(Trace of a scalar is the scalar itself)} \\
&= \text{Trace}(\boldsymbol{\Sigma}_Z) + \|\boldsymbol{\mu}_Z\|^2
\end{aligned}$$

Both forms are useful. □

Proof 2: Unbiasedness of $\hat{\sigma}^2$ using Lemma. We want to compute $\mathbb{E}[\|\boldsymbol{e}\|^2]$. We apply the lemma with $\boldsymbol{Z} = \boldsymbol{e}$.

$$\mathbb{E}[\|\boldsymbol{e}\|^2] = \text{Trace}(\text{Var}(\boldsymbol{e})) + \|\mathbb{E}[\boldsymbol{e}]\|^2$$

We need to find the mean and variance of the residual vector $\boldsymbol{e}$.

Mean of $\boldsymbol{e}$:

$$\mathbb{E}[\boldsymbol{e}] = \mathbb{E}[\mathbf{Q}\boldsymbol{\epsilon}] = \mathbf{Q}\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{Q}\mathbf{0} = \mathbf{0}$$

So, $\|\mathbb{E}[\boldsymbol{e}]\|^2 = 0$.

Variance of $\boldsymbol{e}$:

$$\text{Var}(\boldsymbol{e}) = \text{Var}(\mathbf{Q}\boldsymbol{\epsilon}) = \mathbf{Q}\text{Var}(\boldsymbol{\epsilon})\mathbf{Q}^T \quad \text{(Butterfly formula)}$$

Since $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$ and $\mathbf{Q}$ is symmetric ($\mathbf{Q}^T = \mathbf{Q}$):

$$\text{Var}(\boldsymbol{e}) = \mathbf{Q}(\sigma^2\mathbf{I}_n)\mathbf{Q} = \sigma^2\mathbf{Q}\mathbf{I}_n\mathbf{Q} = \sigma^2\mathbf{Q}^2$$

Since $\mathbf{Q}$ is idempotent ($\mathbf{Q}^2 = \mathbf{Q}$):

$$\text{Var}(\boldsymbol{e}) = \sigma^2\mathbf{Q}$$

Applying the Lemma:

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{e}\|^2] &= \text{Trace}(\text{Var}(\boldsymbol{e})) + \|\mathbb{E}[\boldsymbol{e}]\|^2 \\ &= \text{Trace}(\sigma^2\mathbf{Q}) + 0 \\ &= \sigma^2\text{Trace}(\mathbf{Q})\end{aligned}$$

As established in Proof 1, $\text{Trace}(\mathbf{Q}) = n - p - 1$.

$$\mathbb{E}[\|\boldsymbol{e}\|^2] = \sigma^2(n - p - 1)$$

Therefore,

$$\mathbb{E}[\hat{\sigma}^2] = \frac{\mathbb{E}[\|\boldsymbol{e}\|^2]}{n - p - 1} = \frac{\sigma^2(n - p - 1)}{n - p - 1} = \sigma^2$$

This provides an alternative, perhaps slightly more abstract, route to the same result, leveraging the general lemma about the expected squared norm. □

6 Summary and Next Steps

Let's briefly summarize where we stand:

- We are working within the **Linear Model**: $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$. The parameters $\boldsymbol{\beta}$ and σ^2 are unknown, and $\mathbf{X}$ is considered fixed.
- The **Least Squares Estimator** for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$.
- The **Residual Vector** is $\boldsymbol{e} = \boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{P}_M)\boldsymbol{Y} = \mathbf{Q}\boldsymbol{\epsilon}$.
- The **Estimator for σ^2** is $\hat{\sigma}^2 = S^2 = \frac{\|\boldsymbol{e}\|^2}{n-p-1}$.
- We have shown that both $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are **unbiased estimators** for their respective parameters under the model assumptions:
  - $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
  - $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$
- We also derived the **covariance matrix of the LSE**:

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

We have established the foundational properties of the standard estimators in the linear model framework. In the upcoming lectures, we will delve deeper into the properties of $\hat{\boldsymbol{\beta}}$, culminating in the famous **Gauss-Markov Theorem**. This theorem establishes that, in a specific sense, the LSE is the "best" linear unbiased estimator for $\boldsymbol{\beta}$. This provides a strong theoretical justification for using least squares under the assumptions we have made.