Linear Models: Estimation and Inference under Normality

Lecture Notes Supplement

Your Name/Course Name Here

April 27, 2025

1 Estimating Linear Combinations of Coefficients

In our study of the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we've focused on estimating the entire vector of coefficients $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. Often, however, we are interested in estimating a specific linear combination of these coefficients.

Definition 1.1 (Linear Combination)**.** Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ be the vector of regression coefficients, and let $\boldsymbol{a} = (a_0, a_1, \ldots, a_p)^\top \in \mathbb{R}^{p+1}$ be a fixed, known vector of constants. The scalar quantity

$$\theta := \boldsymbol{a}^\top \boldsymbol{\beta} = \sum_{j=0}^{p} a_j \beta_j \tag{1}$$

is called a **linear combination** of the elements of $\boldsymbol{\beta}$.

Remark 1.2 (Motivation). Why study linear combinations?

- **Individual Coefficients:** Estimating β_j itself corresponds to choosing $\boldsymbol{a}$ with $a_j = 1$ and all other $a_k = 0$.
- **Differences/Contrasts:** Comparing two coefficients, say $\beta_1 - \beta_2$, corresponds to $a_1 = 1, a_2 = -1$, and others zero.
- **Predictions:** For a new set of predictor values $\boldsymbol{x}_* = (1, x_{*1}, \ldots, x_{*p})^\top$, the expected response is $\mathbb{E}[Y_*] = \boldsymbol{x}_*^\top \boldsymbol{\beta}$, which is a linear combination with $\boldsymbol{a} = \boldsymbol{x}_*$.

Given the least squares (LS) estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ for $\boldsymbol{\beta}$, a natural estimator for $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$ is obtained by simply plugging in $\hat{\boldsymbol{\beta}}$:

Definition 1.3 (LS Estimator of a Linear Combination)**.** The LS estimator for $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$ is

$$\hat{\theta} := \boldsymbol{a}^\top \hat{\boldsymbol{\beta}} = \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}. \tag{2}$$

Notice that $\hat{\theta}$ is a linear function of the observation vector $\boldsymbol{Y}$. We can write it as $\hat{\theta} = \boldsymbol{c}^\top \boldsymbol{Y}$, where the vector $\boldsymbol{c} \in \mathbb{R}^n$ is given by

$$\boldsymbol{c} := \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a}. \tag{3}$$

It's worth noting the dimensions: $\boldsymbol{a}$ is in $\mathbb{R}^{p+1}$, while $\boldsymbol{c}$ is in $\mathbb{R}^n$.

Let's examine the properties of this estimator under the standard **linear model assumptions**:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where } \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0} \text{ and } \mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n. \tag{4}$$

These assumptions imply $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$ and $\mathrm{Cov}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}_n$.

Mean of $\hat{\theta}$: Using the linearity of expectation and the unbiasedness of $\hat{\boldsymbol{\beta}}$ ($\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$), we have:

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}] = \boldsymbol{a}^\top \mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{a}^\top \boldsymbol{\beta} = \theta.$$

Thus, $\hat{\theta}$ is an **unbiased** estimator for θ.

Variance of $\hat{\theta}$: Using the linearity property of variance for vector transformations ($V(\boldsymbol{AY}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{A}^\top$) or directly for the linear combination $\hat{\theta} = \boldsymbol{c}^\top \boldsymbol{Y}$:

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}) &= \mathrm{Var}(\boldsymbol{c}^\top \boldsymbol{Y}) \\
&= \boldsymbol{c}^\top \mathrm{Cov}(\boldsymbol{Y})\boldsymbol{c} \\
&= \boldsymbol{c}^\top (\sigma^2 \boldsymbol{I}_n)\boldsymbol{c} \\
&= \sigma^2 \boldsymbol{c}^\top \boldsymbol{c}.
\end{aligned}$$

We can express this variance in terms of $\boldsymbol{a}$ as well:

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}) = \sigma^2 \boldsymbol{c}^\top \boldsymbol{c} &= \sigma^2 \left(\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{a}\right)^\top \left(\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{a}\right) \\
&= \sigma^2 \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a} \\
&= \sigma^2 \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a}.
\end{aligned}$$

So, $\hat{\theta}$ is a linear, unbiased estimator with variance $\sigma^2 \boldsymbol{c}^\top \boldsymbol{c} = \sigma^2 \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a}$.

Is $\hat{\theta}$ the "Best" Linear Unbiased Estimator? We found *an* unbiased linear estimator $\hat{\theta}$. But could there be another estimator, say $\tilde{\theta}$, which is also linear in $\boldsymbol{Y}$ and unbiased for θ, but somehow "better" than $\hat{\theta}$? First, we need to define what "better" means. A standard criterion is the Mean Squared Error (MSE).

Definition 1.4 (Mean Squared Error (MSE))**.** The MSE of an estimator $\hat{\phi}$ for a parameter ϕ is

$$\mathrm{MSE}(\hat{\phi}) := \mathbb{E}_\phi \left[(\hat{\phi} - \phi)^2\right],$$

where the expectation is taken with respect to the distribution governed by the true parameter ϕ. Note that the MSE generally depends on the true value of ϕ.

We say an estimator $\hat{\phi}$ is **better** than another estimator $\tilde{\phi}$ if $\mathrm{MSE}(\hat{\phi}) \leq \mathrm{MSE}(\tilde{\phi})$ for all possible values of the parameter ϕ, with strict inequality for at least one value.

Bias-Variance Decomposition of MSE: A fundamental property of MSE is its decomposition into variance and squared bias. For any estimator $\hat{\phi}$ of ϕ:

$$\begin{aligned}
\mathrm{MSE}(\hat{\phi}) &= \mathbb{E}\left[(\hat{\phi} - \phi)^2\right] \\
&= \mathbb{E}\left[((\hat{\phi} - \mathbb{E}[\hat{\phi}]) + (\mathbb{E}[\hat{\phi}] - \phi))^2\right] \\
&= \mathbb{E}\left[(\hat{\phi} - \mathbb{E}[\hat{\phi}])^2\right] + \mathbb{E}\left[(\mathbb{E}[\hat{\phi}] - \phi)^2\right] + 2\mathbb{E}\left[(\hat{\phi} - \mathbb{E}[\hat{\phi}])(\mathbb{E}[\hat{\phi}] - \phi)\right] \\
&= \mathbb{E}\left[(\hat{\phi} - \mathbb{E}[\hat{\phi}])^2\right] + (\mathbb{E}[\hat{\phi}] - \phi)^2 + 2(\mathbb{E}[\hat{\phi}] - \phi)\underbrace{\mathbb{E}[\hat{\phi} - \mathbb{E}[\hat{\phi}]]}_{=0} \\
&= \underbrace{\mathbb{E}\left[(\hat{\phi} - \mathbb{E}[\hat{\phi}])^2\right]}_{\mathrm{Var}(\hat{\phi})} + \underbrace{(\mathbb{E}[\hat{\phi}] - \phi)^2}_{(\mathrm{bias}(\hat{\phi}))^2}
\end{aligned} \tag{5}$$

where $\text{bias}(\hat{\phi}) = \mathbb{E}[\hat{\phi}] - \phi$.

Crucially, for an **unbiased** estimator, the bias term is zero, so $\text{bias}(\hat{\phi}) = 0$. In this case, the decomposition simplifies beautifully:

$$\text{MSE}(\hat{\phi}) = \text{Var}(\hat{\phi}) \quad (\text{if } \hat{\phi} \text{ is unbiased}).$$

Therefore, if we restrict our search to *unbiased* estimators, finding the "best" estimator in terms of MSE is equivalent to finding the estimator with the *minimum variance*.

This leads us to a central question: Among all linear unbiased estimators for $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$, does the LS estimator $\hat{\theta} = \boldsymbol{a}^\top \hat{\boldsymbol{\beta}}$ have the smallest variance? The answer is yes, as formalized by the famous Gauss-Markov Theorem.

2 The Gauss-Markov Theorem

This theorem is a cornerstone of linear regression theory. It provides a powerful justification for using the least squares estimator, relying only on the assumptions about the first and second moments of the errors, not on any specific distribution like the normal distribution.

Theorem 2.1 (Gauss-Markov)**.** *Assume the linear model holds:* $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ *with* $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$ *and* $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$*. Let* $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$ *be any linear combination of the coefficients. Let* $\hat{\theta} = \boldsymbol{a}^\top \hat{\boldsymbol{\beta}}$ *be the least squares estimator of* θ*, where* $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$*.*

Consider any other estimator $\tilde{\theta}$ *that is also* ***linear*** *in* $\boldsymbol{Y}$ *(i.e.,* $\tilde{\theta} = \boldsymbol{d}^\top \boldsymbol{Y}$ *for some vector* $\boldsymbol{d} \in \mathbb{R}^n$*) and* ***unbiased*** *for* θ *(i.e.,* $\mathbb{E}[\tilde{\theta}] = \theta$ *for all possible* $\boldsymbol{\beta}$*).*

Then, the variance of the LS estimator is less than or equal to the variance of any other such linear unbiased estimator:

$$\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta}) \quad \textit{for all } \boldsymbol{\beta}.$$

For this reason, $\hat{\theta}$ *is called the* ***Best Linear Unbiased Estimator (BLUE)*** *of* θ*.*

Proof. Let $\hat{\theta} = \boldsymbol{c}^\top \boldsymbol{Y}$ where $\boldsymbol{c} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a}$, as defined earlier. Let $\tilde{\theta} = \boldsymbol{d}^\top \boldsymbol{Y}$ be any other linear unbiased estimator for θ.

We can write the vector $\boldsymbol{d}$ in terms of $\boldsymbol{c}$ and a difference vector $\boldsymbol{\Delta}$:

$$\boldsymbol{d} = \boldsymbol{c} + \boldsymbol{\Delta}, \quad \text{where } \boldsymbol{\Delta} = \boldsymbol{d} - \boldsymbol{c} \in \mathbb{R}^n.$$

Now, let's use the unbiasedness condition for $\tilde{\theta}$: $\mathbb{E}[\tilde{\theta}] = \theta$ must hold for all $\boldsymbol{\beta}$.

$$\begin{aligned}
\theta = \mathbb{E}[\tilde{\theta}] &= \mathbb{E}[\boldsymbol{d}^\top \boldsymbol{Y}] \\
&= \mathbb{E}[(\boldsymbol{c} + \boldsymbol{\Delta})^\top \boldsymbol{Y}] \\
&= \mathbb{E}[\boldsymbol{c}^\top \boldsymbol{Y}] + \mathbb{E}[\boldsymbol{\Delta}^\top \boldsymbol{Y}] \\
&= \mathbb{E}[\hat{\theta}] + \boldsymbol{\Delta}^\top \mathbb{E}[\boldsymbol{Y}] \quad (\text{since } \boldsymbol{\Delta} \text{ is constant}) \\
&= \theta + \boldsymbol{\Delta}^\top (\boldsymbol{X}\boldsymbol{\beta}) \quad (\text{since } \hat{\theta} \text{ is unbiased and } \mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta})
\end{aligned}$$

Comparing the start and end of this chain of equalities, we must have:

$$\boldsymbol{\Delta}^\top \boldsymbol{X}\boldsymbol{\beta} = 0 \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^{p+1}.$$

The only way this can hold for all possible vectors $\boldsymbol{\beta}$ is if the vector multiplying $\boldsymbol{\beta}$ is the zero vector. That is,

$$\boldsymbol{\Delta}^\top \boldsymbol{X} = \boldsymbol{0}^\top. \tag{6}$$

This condition captures the implication of unbiasedness for the difference vector $\boldsymbol{\Delta}$.

Now, let's use this condition to examine the relationship between $\boldsymbol{\Delta}$ and $\boldsymbol{c}$. Recall $\boldsymbol{c} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}$. Consider the inner product $\boldsymbol{\Delta}^\top\boldsymbol{c}$:

$$\boldsymbol{\Delta}^\top\boldsymbol{c} = \boldsymbol{\Delta}^\top\left(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}\right) = \underbrace{(\boldsymbol{\Delta}^\top\boldsymbol{X})}_{=\boldsymbol{0}^\top}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a} = \boldsymbol{0}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a} = 0.$$

So, the vectors $\boldsymbol{\Delta}$ and $\boldsymbol{c}$ are orthogonal: $\boldsymbol{\Delta}^\top\boldsymbol{c} = 0$.

Finally, let's calculate the variance of $\tilde{\theta}$:

$$\begin{aligned}
\mathrm{Var}(\tilde{\theta}) &= \mathrm{Var}(\boldsymbol{d}^\top\boldsymbol{Y}) \\
&= \boldsymbol{d}^\top\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{d} \\
&= \boldsymbol{d}^\top(\sigma^2\boldsymbol{I}_n)\boldsymbol{d} \\
&= \sigma^2\boldsymbol{d}^\top\boldsymbol{d} \\
&= \sigma^2(\boldsymbol{c}+\boldsymbol{\Delta})^\top(\boldsymbol{c}+\boldsymbol{\Delta}) \\
&= \sigma^2(\boldsymbol{c}^\top\boldsymbol{c}+\boldsymbol{c}^\top\boldsymbol{\Delta}+\boldsymbol{\Delta}^\top\boldsymbol{c}+\boldsymbol{\Delta}^\top\boldsymbol{\Delta}) \\
&= \sigma^2(\boldsymbol{c}^\top\boldsymbol{c}+0+0+\boldsymbol{\Delta}^\top\boldsymbol{\Delta}) \quad (\text{since } \boldsymbol{\Delta}^\top\boldsymbol{c} = \boldsymbol{c}^\top\boldsymbol{\Delta} = 0) \\
&= \sigma^2(\boldsymbol{c}^\top\boldsymbol{c}+\boldsymbol{\Delta}^\top\boldsymbol{\Delta})
\end{aligned}$$

We know that $\mathrm{Var}(\hat{\theta}) = \sigma^2\boldsymbol{c}^\top\boldsymbol{c}$. Therefore,

$$\mathrm{Var}(\tilde{\theta}) = \mathrm{Var}(\hat{\theta}) + \sigma^2\boldsymbol{\Delta}^\top\boldsymbol{\Delta}.$$

Since $\boldsymbol{\Delta}^\top\boldsymbol{\Delta} = \|\boldsymbol{\Delta}\|^2$ is the squared Euclidean norm of $\boldsymbol{\Delta}$, it must be non-negative ($\boldsymbol{\Delta}^\top\boldsymbol{\Delta} \geq 0$). Thus,

$$\mathrm{Var}(\tilde{\theta}) \geq \mathrm{Var}(\hat{\theta}).$$

Equality holds if and only if $\boldsymbol{\Delta}^\top\boldsymbol{\Delta} = 0$, which implies $\boldsymbol{\Delta} = \boldsymbol{0}$, meaning $\boldsymbol{d} = \boldsymbol{c}$. This confirms that the LS estimator $\hat{\theta}$ has the minimum variance among all linear unbiased estimators. □

Remark 2.2. The Gauss-Markov theorem is truly remarkable. It tells us that without making any assumptions about the *shape* of the error distribution (like normality), as long as the errors are uncorrelated and have constant variance, the least squares procedure naturally yields estimators that are optimal in the class of linear unbiased estimators, judged by the criterion of minimum variance (or equivalently, minimum MSE within this class).

3 Toward Inference: The Need for Distributional Assumptions

We have established the optimality of LS estimators in terms of minimum variance among linear unbiased estimators (point estimation). However, many statistical tasks go beyond point estimation. For example, we often want to:

- Construct **confidence intervals** for coefficients β_j or linear combinations $\theta = \boldsymbol{a}^\top\boldsymbol{\beta}$.
- Perform **hypothesis tests**, such as testing if a particular coefficient is zero ($H_0 : \beta_j = 0$).

To perform these tasks, we need more than just the mean and variance of our estimators; we need to know their *sampling distribution*. This typically requires making stronger assumptions about the distribution of the error term $\boldsymbol{\epsilon}$. The most common and mathematically tractable assumption is that the errors follow a multivariate normal distribution.

Before diving into the normal linear model, let's briefly review some key concepts about multivariate distributions, focusing on the multivariate normal.

3.1 Review of Multivariate Distributions

Consider a random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^\top$.

Joint CDF: The joint cumulative distribution function (CDF) is defined as

$$F_{\boldsymbol{Z}}(z_1, \ldots, z_k) := P(Z_1 \leq z_1, \ldots, Z_k \leq z_k).$$

The CDF always exists and uniquely determines the distribution of $\boldsymbol{Z}$.

Independence: The components $Z_1, \ldots, Z_k$ are mutually (statistically) independent if and only if the joint CDF factors into the product of the marginal CDFs:

$$F_{\boldsymbol{Z}}(z_1, \ldots, z_k) = P(Z_1 \leq z_1) \cdots P(Z_k \leq z_k) = F_{Z_1}(z_1) \cdots F_{Z_k}(z_k)$$

for all $z_1, \ldots, z_k \in \mathbb{R}$.

Joint PDF: If the joint CDF is sufficiently smooth, we can define the joint probability density function (PDF) via differentiation:

$$f_{\boldsymbol{Z}}(z_1, \ldots, z_k) = \frac{\partial^k}{\partial z_1 \cdots \partial z_k} F_{\boldsymbol{Z}}(z_1, \ldots, z_k).$$

When the PDF exists, it provides an equivalent characterization of the distribution, satisfying

$$F_{\boldsymbol{Z}}(z_1, \ldots, z_k) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_k} f_{\boldsymbol{Z}}(u_1, \ldots, u_k) du_k \cdots du_1,$$

and $f_{\boldsymbol{Z}}(\boldsymbol{z}) \geq 0$ with $\int_{\mathbb{R}^k} f_{\boldsymbol{Z}}(\boldsymbol{z}) d\boldsymbol{z} = 1$. For independent continuous random variables, the joint PDF is the product of the marginal PDFs: $f_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{Z_1}(z_1) \cdots f_{Z_k}(z_k)$.

3.2 The Multivariate Normal Distribution

The multivariate normal (MVN) distribution is arguably the most important multivariate distribution in statistics, partly due to the Central Limit Theorem and its mathematical tractability.

Definition 3.1 (Multivariate Normal Distribution)**.** A random vector $\boldsymbol{W} = (W_1, \ldots, W_k)^\top$ is said to have a **multivariate normal distribution** if it can be represented as

$$\boldsymbol{W} \stackrel{d}{=} \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z} \tag{7}$$

where:

- $\boldsymbol{\mu} \in \mathbb{R}^k$ is a constant vector (the mean).

- $\boldsymbol{A} \in \mathbb{R}^{k \times l}$ is a constant matrix.
- $\boldsymbol{Z} = (Z_1, \ldots, Z_l)^\top$ is a random vector whose components Z_i are independent and identically distributed (i.i.d.) standard normal random variables, $Z_i \sim \mathcal{N}(0, 1)$.

The symbol $\stackrel{d}{=}$ means "equal in distribution".

Properties of the Multivariate Normal Distribution:

1. **Mean and Covariance:** If $\boldsymbol{W} \stackrel{d}{=} \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}$ as in Definition 3.1, we can calculate its mean vector and covariance matrix:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{W}] &= \mathbb{E}[\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}] = \boldsymbol{\mu} + \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{0} = \boldsymbol{\mu}. \\
\mathrm{Cov}(\boldsymbol{W}) &= \mathrm{Cov}(\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}) = \mathrm{Cov}(\boldsymbol{A}\boldsymbol{Z}) \\
&= \boldsymbol{A}\mathrm{Cov}(\boldsymbol{Z})\boldsymbol{A}^\top \\
&= \boldsymbol{A}(\boldsymbol{I}_l)\boldsymbol{A}^\top \quad \text{(since } Z_i \text{ are i.i.d. } \mathcal{N}(0,1)) \\
&= \boldsymbol{A}\boldsymbol{A}^\top.
\end{aligned}$$

   Let $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{A}^\top$. Then $\boldsymbol{V}$ is the covariance matrix of $\boldsymbol{W}$. Note that $\boldsymbol{V}$ is always symmetric and positive semi-definite. The representation $\boldsymbol{W} \stackrel{d}{=} \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}$ implies $\mathbb{E}[\boldsymbol{W}] = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{W}) = \boldsymbol{V} = \boldsymbol{A}\boldsymbol{A}^\top$. The mean vector $\boldsymbol{\mu}$ is unique, but the matrix $\boldsymbol{A}$ is not unique for a given $\boldsymbol{V}$ (e.g., if $\boldsymbol{U}$ is an $l \times l$ orthogonal matrix, $\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{U}$ gives $\boldsymbol{A}'(\boldsymbol{A}')^\top = \boldsymbol{A}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{V}$).

2. **Density Function (when $\boldsymbol{V}$ is invertible):** If $l = k$ and the matrix $\boldsymbol{A}_{k \times k}$ is invertible (equivalently, has linearly independent columns), then the covariance matrix $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{A}^\top$ is positive definite (and thus invertible). In this case, the random vector $\boldsymbol{W}$ has a probability density function (PDF) given by:

$$f_{\boldsymbol{W}}(\boldsymbol{w}) = (2\pi)^{-k/2} |\det(\boldsymbol{V})|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu})^\top \boldsymbol{V}^{-1}(\boldsymbol{w} - \boldsymbol{\mu})\right), \quad \boldsymbol{w} \in \mathbb{R}^k.$$

   If $\boldsymbol{V}$ is singular (not invertible), the distribution does not have a density with respect to the Lebesgue measure on $\mathbb{R}^k$; the distribution is concentrated on a lower-dimensional affine subspace.

3. **Notation:** Regardless of whether $\boldsymbol{V}$ is invertible, the distribution of $\boldsymbol{W}$ is completely determined by its mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$. We use the notation

$$\boldsymbol{W} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$$

   to denote that $\boldsymbol{W}$ follows a k-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$.

4. **Linear Combinations are Normal:** If $\boldsymbol{W} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$ and $\boldsymbol{c} \in \mathbb{R}^k$ is a constant vector, then the scalar random variable $Y = \boldsymbol{c}^\top \boldsymbol{W}$ has a univariate normal distribution:

$$\boldsymbol{c}^\top \boldsymbol{W} \sim \mathcal{N}(\boldsymbol{c}^\top \boldsymbol{\mu}, \boldsymbol{c}^\top \boldsymbol{V} \boldsymbol{c}).$$

   This is a crucial property. It follows directly from the definition: if $\boldsymbol{W} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}$, then $\boldsymbol{c}^\top \boldsymbol{W} = \boldsymbol{c}^\top \boldsymbol{\mu} + (\boldsymbol{c}^\top \boldsymbol{A})\boldsymbol{Z}$. Since $(\boldsymbol{c}^\top \boldsymbol{A})\boldsymbol{Z}$ is a linear combination of independent standard

normals, it is itself normally distributed (possibly with variance 0 if $\boldsymbol{c}^\top \boldsymbol{A} = \boldsymbol{0}^\top$). The mean is $\boldsymbol{c}^\top \boldsymbol{\mu}$ and the variance is $(\boldsymbol{c}^\top \boldsymbol{A})(\boldsymbol{c}^\top \boldsymbol{A})^\top = \boldsymbol{c}^\top (\boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{c} = \boldsymbol{c}^\top \boldsymbol{V}\boldsymbol{c}$. *Example:* Taking $\boldsymbol{c} = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ (with 1 in the j-th position) shows that each component W_j is univariate normal:

$$W_j \sim \mathcal{N}(\mu_j, V_{jj}),$$

where V_{jj} is the j-th diagonal element of $\boldsymbol{V}$.

5. **Characterization via Linear Combinations:** The previous property has a converse. A random vector $\boldsymbol{W}$ is multivariate normal if and only if every linear combination $\boldsymbol{c}^\top \boldsymbol{W}$ is univariate normal. Formally:

$$\boldsymbol{W} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{V}) \quad \Longleftrightarrow \quad \boldsymbol{c}^\top \boldsymbol{W} \sim \mathcal{N}(\boldsymbol{c}^\top \boldsymbol{\mu}, \boldsymbol{c}^\top \boldsymbol{V}\boldsymbol{c}) \quad \forall \boldsymbol{c} \in \mathbb{R}^k.$$

(Here $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{W}]$ and $\boldsymbol{V} = \mathrm{Cov}(\boldsymbol{W})$). This provides an alternative way to define or check for multivariate normality.

6. **Linear Transformations Preserve Normality:** If $\boldsymbol{W} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$ and $\boldsymbol{C} \in \mathbb{R}^{m \times k}$ is a constant matrix, then the transformed vector $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{W}$ is also multivariate normal:

$$\boldsymbol{C}\boldsymbol{W} \sim \mathcal{N}_m(\boldsymbol{C}\boldsymbol{\mu}, \boldsymbol{C}\boldsymbol{V}\boldsymbol{C}^\top).$$

This follows because if $\boldsymbol{W} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}$, then $\boldsymbol{C}\boldsymbol{W} = \boldsymbol{C}\boldsymbol{\mu} + (\boldsymbol{C}\boldsymbol{A})\boldsymbol{Z}$. This fits the definition of MVN with mean $\boldsymbol{C}\boldsymbol{\mu}$ and matrix $\boldsymbol{A}' = \boldsymbol{C}\boldsymbol{A}$. The covariance is $\boldsymbol{A}'(\boldsymbol{A}')^\top = (\boldsymbol{C}\boldsymbol{A})(\boldsymbol{C}\boldsymbol{A})^\top = \boldsymbol{C}(\boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{C}^\top = \boldsymbol{C}\boldsymbol{V}\boldsymbol{C}^\top$.

7. **Sums of Independent Normals are Normal:** If $\boldsymbol{W}^{(j)} \sim \mathcal{N}_k(\boldsymbol{\mu}^{(j)}, \boldsymbol{V}^{(j)})$ for $j = 1, \ldots, p$, are independent random vectors, and $d_1, \ldots, d_p$ are scalar constants, then their linear combination is also multivariate normal:

$$\sum_{j=1}^{p} d_j \boldsymbol{W}^{(j)} \sim \mathcal{N}_k\left(\sum_{j=1}^{p} d_j \boldsymbol{\mu}^{(j)}, \sum_{j=1}^{p} d_j^2 \boldsymbol{V}^{(j)}\right).$$

8. **Independence and Zero Covariance:** Let $\boldsymbol{W} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$. Partition $\boldsymbol{W}$ into two subvectors, $\boldsymbol{W}^{(1)}$ and $\boldsymbol{W}^{(2)}$, corresponding to disjoint subsets of indices $\mathcal{I}_1, \mathcal{I}_2 \subseteq \{1, \ldots, k\}$. That is, $\boldsymbol{W}^{(1)} = (W_i : i \in \mathcal{I}_1)$ and $\boldsymbol{W}^{(2)} = (W_j : j \in \mathcal{I}_2)$. A remarkable property of the MVN distribution is that $\boldsymbol{W}^{(1)}$ and $\boldsymbol{W}^{(2)}$ are **statistically independent** if and only if their cross-covariance is zero:

$$\mathrm{Cov}(W_i, W_j) = 0 \quad \forall i \in \mathcal{I}_1, j \in \mathcal{I}_2 \quad \Longleftrightarrow \quad \boldsymbol{W}^{(1)} \text{ and } \boldsymbol{W}^{(2)} \text{ are independent.}$$

In general, zero covariance does not imply independence, but it does for jointly normal random variables. This property is extremely useful in linear models.

4 Distributions Related to the Normal

Several important probability distributions arise from transformations of normal random variables. These are fundamental for constructing confidence intervals and test statistics in the context of the normal linear model.

Definition 4.1 (Chi-square (χ^2) Distribution)**.** If $Z_1, Z_2, \ldots, Z_k$ are independent and identically distributed (i.i.d.) standard normal random variables, $Z_i \sim \mathcal{N}(0,1)$, then the distribution of the sum of their squares,

$$Q = \sum_{j=1}^{k} Z_j^2,$$

is called the **Chi-square distribution** with k degrees of freedom. We denote this by $Q \sim \chi_k^2$. (In R: 'pchisq()', 'qchisq()', 'rchisq()', 'dchisq()').

Proposition 4.2 (Expected Value of χ_k^2)**.** *If $Q \sim \chi_k^2$, then $\mathbb{E}[Q] = k$.*

Proof. Let $Q = \sum_{i=1}^{k} Z_i^2$ where $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$. By linearity of expectation:

$$\mathbb{E}[Q] = \mathbb{E}\left[\sum_{i=1}^{k} Z_i^2\right] = \sum_{i=1}^{k} \mathbb{E}[Z_i^2].$$

For a standard normal random variable Z_i, we know $\mathbb{E}[Z_i] = 0$ and $\mathrm{Var}(Z_i) = 1$. Since $\mathrm{Var}(Z_i) = \mathbb{E}[Z_i^2] - (\mathbb{E}[Z_i])^2$, we have

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

Therefore, $\mathbb{E}[Q] = \sum_{i=1}^{k} 1 = k$. □

Proposition 4.3 (Quadratic Forms in Normal Variables)**.** *Let $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$ (a vector of n i.i.d. standard normal variables). Let $\boldsymbol{P}$ be an $n \times n$ symmetric ($\boldsymbol{P}^\top = \boldsymbol{P}$) and idempotent ($\boldsymbol{P}^2 = \boldsymbol{P}$) matrix with $\mathrm{rank}(\boldsymbol{P}) = r$. Then the quadratic form $\boldsymbol{Z}^\top \boldsymbol{P} \boldsymbol{Z}$ follows a Chi-square distribution with r degrees of freedom:*

$$\boldsymbol{Z}^\top \boldsymbol{P} \boldsymbol{Z} = \|\boldsymbol{P}\boldsymbol{Z}\|^2 \sim \chi_r^2.$$

Proof Sketch. Since $\boldsymbol{P}$ is symmetric and idempotent, it represents an orthogonal projection onto a subspace of dimension $r = \mathrm{rank}(\boldsymbol{P}) = \mathrm{tr}(\boldsymbol{P})$. Such a matrix has eigenvalues that are either 0 or 1, with exactly r eigenvalues equal to 1. By the spectral theorem, $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$ where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{D}$ is diagonal with r ones and $n-r$ zeros. Let $\boldsymbol{W} = \boldsymbol{U}^\top \boldsymbol{Z}$. Since $\boldsymbol{U}^\top$ is orthogonal, $\boldsymbol{W} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{U}^\top \boldsymbol{I}_n \boldsymbol{U}) = \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$. Then

$$\boldsymbol{Z}^\top \boldsymbol{P} \boldsymbol{Z} = \boldsymbol{Z}^\top \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top \boldsymbol{Z} = (\boldsymbol{U}^\top \boldsymbol{Z})^\top \boldsymbol{D} (\boldsymbol{U}^\top \boldsymbol{Z}) = \boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W} = \sum_{i=1}^{n} D_{ii} W_i^2 = \sum_{i=1}^{r} W_i^2,$$

where the sum is taken over the r indices i for which $D_{ii} = 1$. Since W_i are i.i.d. $\mathcal{N}(0,1)$, this sum is, by definition, a χ_r^2 random variable. (The original note uses $\mathbb{E}\|\boldsymbol{P}\boldsymbol{Z}\|^2 = \mathrm{tr}(\mathrm{Cov}[\boldsymbol{P}\boldsymbol{Z}]) = \mathrm{tr}(\boldsymbol{P}\boldsymbol{I}\boldsymbol{P}^\top) = \mathrm{tr}(\boldsymbol{P}) = r$. While this correctly calculates the expected value, it doesn't fully prove the distribution is χ_r^2. The spectral decomposition argument is more complete). □

Definition 4.4 (Student's t-distribution)**.** If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi_k^2$ are independent random variables, then the distribution of the ratio

$$T = \frac{Z}{\sqrt{V/k}}$$

is called the ***t*-distribution** with k degrees of freedom. We denote this by $T \sim t_k$. (In R: 'pt()', 'qt()', 'rt()', 'dt()').

Definition 4.5 (F-distribution)**.** If $V_1 \sim \chi^2_{k_1}$ and $V_2 \sim \chi^2_{k_2}$ are independent random variables, then the distribution of the ratio of scaled Chi-square variables

$$F = \frac{V_1/k_1}{V_2/k_2}$$

is called the **F-distribution** with k_1 (numerator) and k_2 (denominator) degrees of freedom. We denote this by $F \sim \mathrm{F}_{k_1,k_2}$. (In R: 'pf()', 'qf()', 'rf()', 'df()').

5 Inference Under the Normal Linear Model

We now combine the linear model structure with the assumption of normally distributed errors. This allows us to derive exact sampling distributions for our estimators, paving the way for confidence intervals and hypothesis tests.

Recall the standard linear model assumptions (4):

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}, \quad \mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n.$$

We now add the assumption that the errors are multivariate normal.

Definition 5.1 (Normal Linear Model)**.** The **normal linear model** assumes:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n). \tag{8}$$

This implies that the observations $\boldsymbol{Y}$ are also multivariate normal:

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n).$$

Under this model, we can determine the exact distributions of the LS estimator $\hat{\boldsymbol{\beta}}$ and the variance estimator $\hat{\sigma}^2$.

5.1 Distribution of the LS Estimator $\hat{\boldsymbol{\beta}}$

Recall the formula for the LS estimator:

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

Let $\boldsymbol{A} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$, which is a $(p+1) \times n$ matrix. We can write $\hat{\boldsymbol{\beta}}$ in terms of the error vector $\boldsymbol{\epsilon}$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \boldsymbol{A}\boldsymbol{Y} \\
&= \boldsymbol{A}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\epsilon} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\epsilon} \\
&= \boldsymbol{I}_{p+1}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\epsilon}.
\end{aligned}$$

Since $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, and $\hat{\boldsymbol{\beta}}$ is a linear transformation of $\boldsymbol{\epsilon}$ shifted by a constant vector $\boldsymbol{\beta}$, it follows from Property 6 of MVN distributions that $\hat{\boldsymbol{\beta}}$ is also multivariate normal.

We already know its mean and covariance matrix from our earlier work (which did not require normality):

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

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \text{Cov}(\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\epsilon}) = \text{Cov}(\boldsymbol{A}\boldsymbol{\epsilon}) = \boldsymbol{A}\text{Cov}(\boldsymbol{\epsilon})\boldsymbol{A}^\top = \boldsymbol{A}(\sigma^2\boldsymbol{I}_n)\boldsymbol{A}^\top = \sigma^2\boldsymbol{A}\boldsymbol{A}^\top.$$

Calculating $\boldsymbol{A}\boldsymbol{A}^\top$:

$$\boldsymbol{A}\boldsymbol{A}^\top = \left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right)\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right)^\top = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}.$$

So, $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.

Combining these results, we have the sampling distribution of the LS estimator under the normal linear model:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right). \tag{9}$$

5.2 Distribution of the Variance Estimator $\hat{\sigma}^2$

Recall the definition of the unbiased variance estimator:

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{e}\|^2}{n-p-1} = \frac{SSE}{n-p-1},$$

where $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{P_X}\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{P_X})\boldsymbol{Y}$ are the residuals, and $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is the projection matrix onto the column space of $\boldsymbol{X}$. Let $\boldsymbol{Q} = \boldsymbol{I}_n - \boldsymbol{P_X}$. Then $\boldsymbol{e} = \boldsymbol{Q}\boldsymbol{Y}$.

We can express the residuals in terms of the errors $\boldsymbol{\epsilon}$:

$$\boldsymbol{e} = \boldsymbol{Q}\boldsymbol{Y} = \boldsymbol{Q}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Q}\boldsymbol{\epsilon}.$$

Since $\boldsymbol{Q}$ projects onto the space orthogonal to the column space of $\boldsymbol{X}$, we have $\boldsymbol{Q}\boldsymbol{X} = \boldsymbol{0}$. Thus,

$$\boldsymbol{e} = \boldsymbol{Q}\boldsymbol{\epsilon}.$$

The sum of squared errors (SSE) is $\|\boldsymbol{e}\|^2 = \|\boldsymbol{Q}\boldsymbol{\epsilon}\|^2 = \boldsymbol{\epsilon}^\top\boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{\epsilon}$. The matrix $\boldsymbol{Q}$ is symmetric ($\boldsymbol{Q}^\top = \boldsymbol{Q}$) and idempotent ($\boldsymbol{Q}^2 = \boldsymbol{Q}$) because it's an orthogonal projection matrix. Its rank is $\text{rank}(\boldsymbol{Q}) = \text{tr}(\boldsymbol{Q}) = \text{tr}(\boldsymbol{I}_n - \boldsymbol{P_X}) = \text{tr}(\boldsymbol{I}_n) - \text{tr}(\boldsymbol{P_X}) = n - \text{tr}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) = n - \text{tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}) = n - \text{tr}(\boldsymbol{I}_{p+1}) = n - (p+1)$.

Now, consider the random vector $\boldsymbol{Z} = \boldsymbol{\epsilon}/\sigma$. Since $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, we have $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$. The SSE can be written as:

$$SSE = \|\boldsymbol{e}\|^2 = \|\boldsymbol{Q}\boldsymbol{\epsilon}\|^2 = \|\sigma\boldsymbol{Q}(\boldsymbol{\epsilon}/\sigma)\|^2 = \sigma^2\|\boldsymbol{Q}\boldsymbol{Z}\|^2.$$

We apply Proposition 4.3 with the matrix $\boldsymbol{P} = \boldsymbol{Q}$ and vector $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$. Since $\boldsymbol{Q}$ is symmetric, idempotent, and has rank $r = n - p - 1$, we conclude that

$$\frac{SSE}{\sigma^2} = \frac{\|\boldsymbol{e}\|^2}{\sigma^2} = \|\boldsymbol{Q}\boldsymbol{Z}\|^2 \sim \chi^2_{n-p-1}.$$

Relating this back to $\hat{\sigma}^2$:

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2_{n-p-1}.$$

This gives us the sampling distribution of the variance estimator $\hat{\sigma}^2$. It follows a scaled Chi-square distribution.

5.3 Joint Distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$

A crucial result for constructing t-statistics and F-statistics is that, under the normal linear model, the LS estimator $\hat{\boldsymbol{\beta}}$ is statistically independent of the variance estimator $\hat{\sigma}^2$.

To show this, we will show that $\hat{\boldsymbol{\beta}}$ is independent of the residual vector $\boldsymbol{e}$, since $\hat{\sigma}^2$ is solely a function of $\boldsymbol{e}$ (specifically, $\|\boldsymbol{e}\|^2$).

We know $\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}$ and $\boldsymbol{e} = \boldsymbol{Q}\boldsymbol{Y}$, where $\boldsymbol{A} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and $\boldsymbol{Q} = \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}$. Consider the joint vector:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \boldsymbol{e} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{Q} \end{pmatrix} \boldsymbol{Y}.$$

Since $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$, this stacked vector is also multivariate normal because it's a linear transformation of $\boldsymbol{Y}$.

Now, we examine the covariance between the blocks $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$:

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) &= \mathrm{Cov}(\boldsymbol{A}\boldsymbol{Y}, \boldsymbol{Q}\boldsymbol{Y}) \\
&= \boldsymbol{A}\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{Q}^\top \\
&= \boldsymbol{A}(\sigma^2\boldsymbol{I}_n)\boldsymbol{Q} \quad \text{(since } \boldsymbol{Q} \text{ is symmetric)} \\
&= \sigma^2\boldsymbol{A}\boldsymbol{Q}
\end{aligned}$$

Let's compute $\boldsymbol{A}\boldsymbol{Q}$:

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{Q} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{I}_{p+1}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top \\
&= \boldsymbol{0}_{(p+1)\times n} \quad \text{(The zero matrix)}
\end{aligned}$$

Therefore, $\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) = \boldsymbol{0}$.

Since $(\hat{\boldsymbol{\beta}}, \boldsymbol{e})$ are jointly multivariate normal and their covariance is zero, we can invoke Property 8 of MVN distributions to conclude that $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ are **statistically independent**.

Because $\hat{\sigma}^2 = \|\boldsymbol{e}\|^2/(n-p-1)$ is a function of $\boldsymbol{e}$ only, it follows that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are also **statistically independent**.

Theorem 5.2 (Independence of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$)**.** *Under the normal linear model* (8)*, the least squares estimator $\hat{\boldsymbol{\beta}}$ is statistically independent of the residual variance estimator $\hat{\sigma}^2$.*

This independence is fundamental. It allows us, for example, to form t-statistics like

$$\frac{(\boldsymbol{a}^\top\hat{\boldsymbol{\beta}} - \boldsymbol{a}^\top\boldsymbol{\beta})}{\sqrt{\hat{\sigma}^2\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}}}$$

where the numerator (related to $\hat{\boldsymbol{\beta}}$) is independent of the denominator (related to $\hat{\sigma}^2$), leading to a t-distribution (after appropriate scaling), which forms the basis for confidence intervals and hypothesis tests for $\boldsymbol{a}^\top\boldsymbol{\beta}$.