Lecture Notes: Inference Under the Linear Model

Your Name/Course Name Here

April 22, 2025

1 From Least Squares Fitting to Statistical Inference

In our exploration of linear models so far, we've focused on the method of **Least Squares (LS)**. Given a design matrix $\boldsymbol{X}$ (size $n \times (p+1)$, where n is the number of observations and p is the number of predictor variables, plus one for an intercept) and a vector of observed responses $\boldsymbol{Y}$ (size $n \times 1$), we found the coefficient vector $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squared differences between the observed responses and the values predicted by the linear model. This LS estimator is given by:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}, \quad \text{where} \quad \boldsymbol{A} := (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$$

Assuming, of course, that $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible, which typically holds if $n \geq p+1$ and the columns of $\boldsymbol{X}$ are linearly independent.

From this estimate, we defined two important vectors:

- The vector of **fitted values**: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. This represents the projection of the observed data $\boldsymbol{Y}$ onto the column space of $\boldsymbol{X}$.
- The vector of **residuals**: $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. This captures the part of the data *not* explained by the linear fit.

A fundamental geometric property we discovered is that the residual vector is orthogonal to the fitted values vector, $\boldsymbol{e} \perp \hat{\boldsymbol{Y}}$, and indeed, orthogonal to any vector in the column space of $\boldsymbol{X}$.

Remark 1.1 (Algebra, Not Statistics (Yet)). It's essential to remember that everything described above is derived from the algebraic goal of minimizing $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$. We haven't needed to make any assumptions about probability distributions or random errors. These results hold for any given dataset $(\boldsymbol{X}, \boldsymbol{Y})$.

Our goal now is to move beyond simply finding the "best fit" line or hyperplane. We want to perform **statistical inference**. This involves asking questions like:

- How uncertain is our estimate $\hat{\boldsymbol{\beta}}$? Can we construct confidence intervals for the true coefficients $\boldsymbol{\beta}$?
- Is a particular predictor variable significantly associated with the response? (i.e., can we test hypotheses like $H_0 : \beta_j = 0$?)
- How much variability in the response is inherent noise versus explained by the model?

To answer these questions, we need to introduce a probabilistic framework – the **linear model assumptions** – which describes how the data $\boldsymbol{Y}$ are generated.

2 The Linear Model: Assumptions and Basic Properties

We now formally adopt the standard (or Gaussian) linear model.

Definition 2.1 (The Linear Model)**.** The linear model assumes that the response vector $\boldsymbol{Y}$ is generated according to:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where:

- $\boldsymbol{X}$ is the $n \times (p+1)$ design matrix, treated as fixed and known.
- $\boldsymbol{\beta}$ is the $(p+1) \times 1$ vector of *true, unknown* population coefficients.
- $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of *unobserved random errors*, satisfying:

(LM1) **Zero Mean**: $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$. (Equivalently, $\mathbb{E}[\epsilon_i] = 0$ for each $i = 1, \ldots, n$).

(LM2) **Constant Variance (Homoscedasticity)**: $\mathrm{Var}(\epsilon_i) = \sigma^2$ for all i, where $\sigma^2 > 0$ is an unknown parameter.

(LM3) **Uncorrelated Errors**: $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$.

Assumptions (LM2) and (LM3) can be compactly written using the covariance matrix of the error vector: $\mathrm{Cov}(\boldsymbol{\epsilon}) = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] - (\mathbb{E}[\boldsymbol{\epsilon}])(\mathbb{E}[\boldsymbol{\epsilon}])^\top = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \sigma^2\boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Often, an additional assumption is made for exact inference:

(LM4) **Normality**: The errors ϵ_i are normally distributed. Combined with (LM1)-(LM3), this means $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$.

For the results in this section (calculating means and covariances), we only need assumptions (LM1)-(LM3). The normality assumption (LM4) will become crucial later when we discuss distributions of test statistics (like t-tests and F-tests).

Under these assumptions, $\boldsymbol{Y}$ becomes a random vector. Let's find its mean and covariance matrix.

Mean of $\boldsymbol{Y}$: Using the linearity of expectation and (LM1):

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

(Note: $\boldsymbol{X}\boldsymbol{\beta}$ is considered a constant vector in this expectation, as $\boldsymbol{X}$ is fixed and $\boldsymbol{\beta}$ is a fixed, albeit unknown, parameter vector).

Covariance of $\boldsymbol{Y}$: Using properties of covariance (adding a constant vector doesn't change covariance) and the definition $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$:

$$\mathrm{Cov}(\boldsymbol{Y}) = \mathrm{Cov}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$$

3 Statistical Properties of the LS Estimator $\hat{\boldsymbol{\beta}}$

Now that $\boldsymbol{Y}$ is a random vector, our LS estimator $\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$ also becomes a random vector. We can investigate its statistical properties, specifically its mean and covariance matrix.

3.1 Mean of $\hat{\boldsymbol{\beta}}$ (Unbiasedness)

Let's calculate the expected value of our estimator.

$$
\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[\boldsymbol{A}\boldsymbol{Y}] \\
&= \boldsymbol{A}\mathbb{E}[\boldsymbol{Y}] \quad \text{(since } \boldsymbol{A} \text{ is a constant matrix)} \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}) \quad \text{(substituting } \mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}\text{)} \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{\beta} \\
&= \boldsymbol{I}_{p+1}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}
$$

This proves a fundamental result:

Proposition 3.1. *Under the linear model assumptions (LM1)-(LM3), the Least Squares estimator $\hat{\boldsymbol{\beta}}$ is an* ***unbiased estimator*** *of the true coefficient vector $\boldsymbol{\beta}$. That is, $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.*

This means that if we were to repeat our experiment many times and calculate $\hat{\boldsymbol{\beta}}$ each time, the average of these estimates would converge to the true value $\boldsymbol{\beta}$. This is a very desirable property for an estimator.

3.2 Covariance Matrix of $\hat{\boldsymbol{\beta}}$

Next, let's find the covariance matrix of $\hat{\boldsymbol{\beta}}$. Recall the property for a constant matrix $\boldsymbol{B}$ and a random vector $\boldsymbol{Z}$: $\mathrm{Cov}(\boldsymbol{B}\boldsymbol{Z}) = \boldsymbol{B}\,\mathrm{Cov}(\boldsymbol{Z})\boldsymbol{B}^\top$. We apply this with $\boldsymbol{Z} = \boldsymbol{Y}$ and $\boldsymbol{B} = \boldsymbol{A} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

$$
\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathrm{Cov}(\boldsymbol{A}\boldsymbol{Y}) \\
&= \boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{A}^\top \\
&= \boldsymbol{A}(\sigma^2\boldsymbol{I}_n)\boldsymbol{A}^\top \quad \text{(substituting } \mathrm{Cov}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n\text{)} \\
&= \sigma^2\boldsymbol{A}\boldsymbol{A}^\top \\
&= \sigma^2\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right]\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right]^\top \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\left((\boldsymbol{X}^\top)^\top((\boldsymbol{X}^\top\boldsymbol{X})^{-1})^\top\right) \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}((\boldsymbol{X}^\top\boldsymbol{X})^{-1}) \quad \text{(since } (\boldsymbol{X}^\top\boldsymbol{X})^{-1} \text{ is symmetric)} \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= \sigma^2\boldsymbol{I}_{p+1}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}
\end{aligned}
$$

We have derived the covariance matrix of the LS estimator:

Proposition 3.2. *Under the linear model assumptions (LM1)-(LM3), the covariance matrix of the LS estimator $\hat{\boldsymbol{\beta}}$ is given by:*

$$\boxed{\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}}$$

Remark 3.3 (Interpreting $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$). This $(p+1)\times(p+1)$ matrix is crucial for inference.

- The diagonal entries give the variances of the individual coefficient estimators: $\mathrm{Var}(\hat{\beta}_j) = \sigma^2[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}]_{jj}$. This measures the precision of each estimate.
- The off-diagonal entries give the covariances between different coefficient estimators: $\mathrm{Cov}(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}]_{jk}$. This tells us how the estimates tend to vary together.
- Notice the dependence on σ^2: higher underlying noise variance leads to higher variance (less precision) in our estimates.
- The term $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ reflects the influence of the experimental design or data structure. For example, multicollinearity (near linear dependence among columns of $\boldsymbol{X}$) tends to inflate the diagonal elements of $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, increasing the variance of the corresponding coefficient estimates.

To use this covariance matrix for practical inference (like constructing confidence intervals or hypothesis tests), we need the value of σ^2. Since σ^2 is typically unknown, we must estimate it from the data.

4 Estimating the Error Variance σ^2

4.1 Motivation and Definition

How can we estimate the underlying noise level σ^2? Intuitively, the residuals $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ represent the discrepancy between our observations and the model's fit. The magnitude of these residuals should reflect the magnitude of the true errors $\boldsymbol{\epsilon}$.

A natural quantity to consider is the **Sum of Squared Residuals (SSR)**:

$$\mathrm{SSR} = \|\boldsymbol{e}\|^2 = \sum_{i=1}^{n} e_i^2$$

Since $\sigma^2 = \mathrm{Var}(\epsilon_i) = \mathbb{E}[\epsilon_i^2]$ (because $\mathbb{E}[\epsilon_i] = 0$), we might think SSR is related to $n\sigma^2$. However, the residuals e_i are not the same as the true errors ϵ_i. The residuals are calculated using the *estimated* coefficients $\hat{\boldsymbol{\beta}}$, which depend on the data $\boldsymbol{Y}$. This process of estimating $\boldsymbol{\beta}$ "uses up" some information from the data.

Specifically, we estimated $p+1$ parameters (the components of $\boldsymbol{\beta}$). It turns out that the appropriate divisor for SSR to get an unbiased estimate of σ^2 is not n, but $n-(p+1)$, the **residual degrees of freedom**.

Definition 4.1 (Unbiased Estimator of σ^2)**.** The unbiased estimator of the error variance σ^2, often denoted by $\hat{\sigma}^2$ or s^2, is defined as:

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

This quantity is also known as the **Residual Mean Square** (RMS or MSE).

We will now rigorously prove that this definition indeed yields an unbiased estimator.

Proposition 4.2. *Under the linear model assumptions (LM1)-(LM3), the estimator $\hat{\sigma}^2$ defined above is an unbiased estimator of σ^2. That is, $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.*

We offer two distinct proofs of this important result. They provide different perspectives and utilize different mathematical tools.

4.2 Proof 1: Using Projection Matrices

This proof relies heavily on the geometric interpretation of least squares in terms of projections.

First Proof of Proposition 4.2. Let $M = \text{Im}(\boldsymbol{X})$ be the column space of the design matrix $\boldsymbol{X}$. This is the subspace of $\mathbb{R}^n$ spanned by the columns of $\boldsymbol{X}$. Assuming $\boldsymbol{X}$ has full column rank, the dimension of M is $p+1$.

Recall that $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\boldsymbol{Y}$ onto M. Let $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ be the $n \times n$ projection matrix onto M. So, $\hat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$.

The residual vector is $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y}$. Let $\boldsymbol{Q} := \boldsymbol{I}_n - \boldsymbol{P}$. $\boldsymbol{Q}$ is the projection matrix onto the orthogonal complement of M, denoted $M^\perp$. We know the following properties of $\boldsymbol{Q}$:

- It is symmetric: $\boldsymbol{Q}^\top = \boldsymbol{Q}$.
- It is idempotent: $\boldsymbol{Q}^2 = \boldsymbol{Q}\boldsymbol{Q} = \boldsymbol{Q}$.
- It annihilates vectors in M: $\boldsymbol{Q}\boldsymbol{X} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{P}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{0}$. (Since $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ because columns of $\boldsymbol{X}$ are in M).

Now, let's express the residual vector in terms of the true errors $\boldsymbol{\epsilon}$. Using the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$:

$$\boldsymbol{e} = \boldsymbol{Q}\boldsymbol{Y} = \boldsymbol{Q}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Q}\boldsymbol{\epsilon} = \boldsymbol{0} \cdot \boldsymbol{\beta} + \boldsymbol{Q}\boldsymbol{\epsilon} = \boldsymbol{Q}\boldsymbol{\epsilon}$$

This crucial step shows that the residual vector $\boldsymbol{e}$ is simply the projection of the true (unobserved) error vector $\boldsymbol{\epsilon}$ onto the subspace $M^\perp$, which is orthogonal to the space spanned by our predictors.

Now we compute the expected value of the Sum of Squared Residuals, $\mathbb{E}[\|\boldsymbol{e}\|^2]$.

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{e}\|^2] &= \mathbb{E}[\|\boldsymbol{Q}\boldsymbol{\epsilon}\|^2] \\
&= \mathbb{E}[(\boldsymbol{Q}\boldsymbol{\epsilon})^\top(\boldsymbol{Q}\boldsymbol{\epsilon})] \\
&= \mathbb{E}[\boldsymbol{\epsilon}^\top\boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{\epsilon}] \\
&= \mathbb{E}[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}] \quad (\text{using } \boldsymbol{Q}^\top = \boldsymbol{Q} \text{ and } \boldsymbol{Q}^2 = \boldsymbol{Q})
\end{aligned}$$

The term $\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}$ is a quadratic form in the random vector $\boldsymbol{\epsilon}$. We can write it explicitly as $\sum_{i=1}^n \sum_{j=1}^n Q_{ij}\epsilon_i\epsilon_j$. By linearity of expectation:

$$\mathbb{E}[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}] = \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n Q_{ij}\epsilon_i\epsilon_j\right] = \sum_{i=1}^n \sum_{j=1}^n Q_{ij}\mathbb{E}[\epsilon_i\epsilon_j]$$

From the linear model assumptions (LM1)-(LM3), we know $\mathbb{E}[\epsilon_i] = 0$ and $\text{Cov}(\epsilon_i, \epsilon_j) = \mathbb{E}[\epsilon_i \epsilon_j]$. Thus:

$$\mathbb{E}[\epsilon_i \epsilon_j] = \begin{cases} \text{Var}(\epsilon_i) = \sigma^2, & \text{if } i = j \\ \text{Cov}(\epsilon_i, \epsilon_j) = 0, & \text{if } i \neq j \end{cases}$$

Substituting this into our sum:

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

The sum $\sum_{i=1}^{n} Q_{ii}$ is precisely the trace of the matrix $\boldsymbol{Q}$, denoted $\text{tr}(\boldsymbol{Q})$. So, we have $\mathbb{E}[\|\boldsymbol{e}\|^2] = \sigma^2 \,\text{tr}(\boldsymbol{Q})$.

What is the trace of the projection matrix $\boldsymbol{Q}$? The trace of any projection matrix is equal to the dimension of the subspace it projects onto. $\boldsymbol{Q}$ projects onto $M^{\perp}$. We know $\dim(\mathbb{R}^n) = n$ and $\dim(M) = p+1$ (assuming $\boldsymbol{X}$ has full rank). By the rank-nullity theorem or properties of orthogonal complements, $\dim(M^{\perp}) = \dim(\mathbb{R}^n) - \dim(M) = n - (p+1)$. Therefore, $\text{rank}(\boldsymbol{Q}) = \dim(M^{\perp}) = n - p - 1$. Since the trace of a projection matrix equals its rank, we have $\text{tr}(\boldsymbol{Q}) = n - p - 1$.

Substituting this back into our expectation calculation:

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

Finally, we can find the expectation of our proposed estimator $\hat{\sigma}^2$:

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

This completes the proof that $\hat{\sigma}^2$ is an unbiased estimator of σ^2. □

Remark 4.3 (Degrees of Freedom). The quantity $n - p - 1$ is often called the **degrees of freedom for error** (or residual degrees of freedom). It represents the number of independent pieces of information in the data that are available for estimating the variance σ^2, after having already used $p + 1$ degrees of freedom to estimate the coefficients in $\boldsymbol{\beta}$.

4.3 Proof 2: Using a General Lemma about Expected Norms

An alternative, perhaps more abstract, proof uses a general result about the expected squared norm of a random vector.

Lemma 4.4. *Let $\boldsymbol{Z}$ be a random vector in $\mathbb{R}^k$ with mean $\boldsymbol{\mu_Z} = \mathbb{E}[\boldsymbol{Z}]$ and covariance matrix $\text{Cov}(\boldsymbol{Z})$. Then,*

$$\mathbb{E}[\|\boldsymbol{Z}\|^2] = \text{tr}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]) = \text{tr}(\text{Cov}(\boldsymbol{Z}) + \boldsymbol{\mu_Z}\boldsymbol{\mu_Z}^\top)$$

As a special case, if the mean is zero ($\boldsymbol{\mu_Z} = \boldsymbol{0}$), then $\mathbb{E}[\|\boldsymbol{Z}\|^2] = \text{tr}(\text{Cov}(\boldsymbol{Z}))$.

Proof of Lemma 4.4. We start with the definition of the squared Euclidean norm $\|\boldsymbol{Z}\|^2 = \boldsymbol{Z}^\top \boldsymbol{Z}$.

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{Z}\|^2] &= \mathbb{E}[\boldsymbol{Z}^\top \boldsymbol{Z}] \\
&\overset{(a)}{=} \mathbb{E}[\operatorname{tr}(\boldsymbol{Z}^\top \boldsymbol{Z})] \quad \text{(Since } \boldsymbol{Z}^\top \boldsymbol{Z} \text{ is a } 1 \times 1 \text{ matrix, its trace is itself)} \\
&\overset{(b)}{=} \mathbb{E}[\operatorname{tr}(\boldsymbol{Z}\boldsymbol{Z}^\top)] \quad \text{(Using the cyclic property of trace: } \operatorname{tr}(AB) = \operatorname{tr}(BA)) \\
&\overset{(c)}{=} \operatorname{tr}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]) \quad \text{(Linearity of trace and expectation allows swapping them)} \\
&\overset{(d)}{=} \operatorname{tr}(\operatorname{Cov}(\boldsymbol{Z}) + \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{\mu}_{\boldsymbol{Z}}^\top) \quad \text{(Using the definition } \operatorname{Cov}(\boldsymbol{Z}) = \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top] - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{\mu}_{\boldsymbol{Z}}^\top)
\end{aligned}$$

This establishes the lemma. The special case follows immediately by setting $\boldsymbol{\mu}_{\boldsymbol{Z}} = \mathbf{0}$. □

Now we apply this lemma to prove Proposition 4.2.

Alternative Proof of Proposition 4.2. We want to compute $\mathbb{E}[\|\boldsymbol{e}\|^2]$. We apply Lemma 4.4 with the random vector $\boldsymbol{Z} = \boldsymbol{e}$. To do this, we first need the mean and covariance matrix of $\boldsymbol{e}$.

Recall from the first proof that $\boldsymbol{e} = \boldsymbol{Q}\boldsymbol{Y}$, where $\boldsymbol{Q} = \boldsymbol{I}_n - \boldsymbol{P}$. The mean of $\boldsymbol{e}$ is:

$$\mathbb{E}[\boldsymbol{e}] = \mathbb{E}[\boldsymbol{Q}\boldsymbol{Y}] = \boldsymbol{Q}\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{Q}(\boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{Q}\boldsymbol{X})\boldsymbol{\beta} = \mathbf{0} \cdot \boldsymbol{\beta} = \mathbf{0}$$

So, the mean vector $\boldsymbol{\mu}_{\boldsymbol{e}}$ is the zero vector.

The covariance matrix of $\boldsymbol{e}$ is:

$$\begin{aligned}
\operatorname{Cov}(\boldsymbol{e}) &= \operatorname{Cov}(\boldsymbol{Q}\boldsymbol{Y}) \\
&= \boldsymbol{Q}\operatorname{Cov}(\boldsymbol{Y})\boldsymbol{Q}^\top \\
&= \boldsymbol{Q}(\sigma^2 \boldsymbol{I}_n)\boldsymbol{Q}^\top \quad \text{(using } \operatorname{Cov}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}_n) \\
&= \sigma^2 \boldsymbol{Q}\boldsymbol{I}_n\boldsymbol{Q}^\top \\
&= \sigma^2 \boldsymbol{Q}\boldsymbol{Q}^\top \\
&= \sigma^2 \boldsymbol{Q}\boldsymbol{Q} \quad \text{(since } \boldsymbol{Q} \text{ is symmetric)} \\
&= \sigma^2 \boldsymbol{Q} \quad \text{(since } \boldsymbol{Q} \text{ is idempotent)}
\end{aligned}$$

So, $\operatorname{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{Q}$.

Now we apply the special case of Lemma 4.4 (since $\boldsymbol{\mu}_{\boldsymbol{e}} = \mathbf{0}$):

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{e}\|^2] &= \operatorname{tr}(\operatorname{Cov}(\boldsymbol{e})) \\
&= \operatorname{tr}(\sigma^2 \boldsymbol{Q}) \\
&= \sigma^2 \operatorname{tr}(\boldsymbol{Q}) \quad \text{(Linearity of trace)}
\end{aligned}$$

As established in the first proof using properties of projection matrices, $\operatorname{tr}(\boldsymbol{Q}) = n - p - 1$. Therefore,

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

Finally, the expectation of our estimator $\hat{\sigma}^2$ is:

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

This provides a second confirmation that $\hat{\sigma}^2$ is an unbiased estimator for σ^2. □

With these results – the unbiasedness of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, and the covariance matrix of $\hat{\boldsymbol{\beta}}$ – we have laid the groundwork for constructing confidence intervals and hypothesis tests concerning the regression coefficients $\boldsymbol{\beta}$, which are central tasks in statistical inference for linear models.