Data Science for Mathematicians

Lesson 2:
Linear Regression from a Geometric Perspective

December 18, 2025

Abstract

This lecture reframes the problem of linear regression from a statistical fitting exercise into a fundamental problem of linear algebra. It establishes that the ordinary least squares solution is geometrically equivalent to finding the orthogonal projection of the response vector onto the subspace spanned by the feature vectors, known as the column space of the data matrix. This geometric constraint of orthogonality is shown to directly and uniquely yield the Normal Equations, providing a coordinate-free, intuitive, and rigorous understanding of the linear regression model. The equivalence between this geometric derivation and the traditional analytical approach of minimizing a squared-error loss function via calculus is then formally proven, revealing a profound unity between two foundational branches of mathematics in the context of data analysis.

Contents

1 From Data Clouds to Best-Fit Subspaces

1.1 Recapitulation of the Variable Space View

In our introductory lecture, we established a foundational perspective for this course: viewing data through a geometric lens. We conceptualized an entire dataset, comprising n observations and p features, as a single mathematical object—the data matrix $X \in \mathbb{R}^{n \times p}$. Our primary interpretation of this matrix was as a collection of n row vectors, $\{\mathbf{x}_i^T\}_{i=1}^n$, where each vector $\mathbf{x}_i \in \mathbb{R}^p$ represents a single observation or data point within a p-dimensional feature space.

This *variable space* or *scatter plot* view is the most common and intuitive way to visualize data, particularly when the number of features is small ($p = 2$ or $p = 3$). In this space, the problem of linear regression is often described as finding a hyperplane of dimension $p - 1$ that *best fits* this cloud of n points. While visually appealing, this perspective becomes abstract in high-dimensional settings and can be algebraically cumbersome for deriving the properties of the model. Today, we perform a pivotal shift in perspective that unlocks a deeper, more powerful understanding of the problem.

1.2 The Observation Space View

The central conceptual leap of this lecture is to reinterpret the geometry of our data matrix. We will now view the same matrix X not as a collection of row vectors in $\mathbb{R}^p$, but as a collection of p column vectors, each residing in an n-dimensional space, $\mathbb{R}^n$:

$$X = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_p \\ | & | & & | \end{bmatrix}$$

Here, each column vector $\mathbf{x}_j \in \mathbb{R}^n$ represents all n observations for the j-th feature. This shift moves our geometric setting from the p-dimensional feature space to an n-dimensional *observation space*, where each dimension corresponds to a specific observation (e.g., a specific patient, a specific day, a specific experimental unit) in our dataset.

This change in viewpoint is not merely a notational convenience; it is a profound reframing of the problem. While the observation space is often of extremely high dimension (since n can be large) and thus difficult to visualize directly, it is algebraically supreme. It allows us to treat the entire set of observations for a given feature as a single vector. This transforms the problem from fitting a shape to individual points into a problem of vector approximation within a single, unified vector space. This perspective allows the powerful machinery of linear algebra—subspaces, orthogonality, and projections—to be brought to bear on the entire dataset simultaneously.

1.3 The Supervised Learning Problem in $\mathbb{R}^n$

Within this new geometric framework, the problem of supervised learning can be stated with remarkable clarity. We are given the p feature vectors $\{\mathbf{x}_j\}_{j=1}^p$, all residing in $\mathbb{R}^n$. We are also given a corresponding target or response vector, $\mathbf{y} \in \mathbb{R}^n$, which contains all n observed outcomes.

The objective of a *linear* model is to find the best possible approximation of the target vector $\mathbf{y}$ using a linear combination of the given feature vectors. That is, we seek a set of scalar coefficients, $\{\beta_1, \beta_2, \ldots, \beta_p\}$, to construct a predicted vector, $\hat{\mathbf{y}}$, such that:

$$\hat{\mathbf{y}} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_p \mathbf{x}_p$$

This equation has a direct geometric interpretation: the predicted vector $\hat{\mathbf{y}}$ is constructed by scaling and adding the feature vectors $\mathbf{x}_j$.

1.4 Projection onto a Subspace

The set of all possible linear combinations of the feature vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ forms a vector subspace of $\mathbb{R}^n$. This subspace is precisely the **column space** of the data matrix X, denoted by

$$\text{Col}(X) = \{X\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\} = \text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}.$$

The dimensionality of this subspace is equal to the rank of the matrix X, which, in a typical regression setting, is equal to the number of linearly independent features, p.

With this definition, the problem of linear regression can be restated in purely geometric terms, devoid of statistical language:

Find the vector $\hat{\mathbf{y}}$ *in the subspace* $\text{Col}(X)$ *that is closest to the target vector* $\mathbf{y}$.

As we know from linear algebra, the unique vector in a subspace that is closest to an external vector is its **orthogonal projection**. Therefore, the geometric goal of linear regression is to find

$$\hat{\mathbf{y}} = \text{proj}_{\text{Col}(X)}\mathbf{y}.$$

Example 1.1. Consider a simple data matrix in $\mathbb{R}^3$ with two feature vectors:

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

The column space $\text{Col}(X)$ consists of all vectors of the form:

$$X\boldsymbol{\beta} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 = \beta_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \beta_2 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_1 + \beta_2 \end{pmatrix}.$$

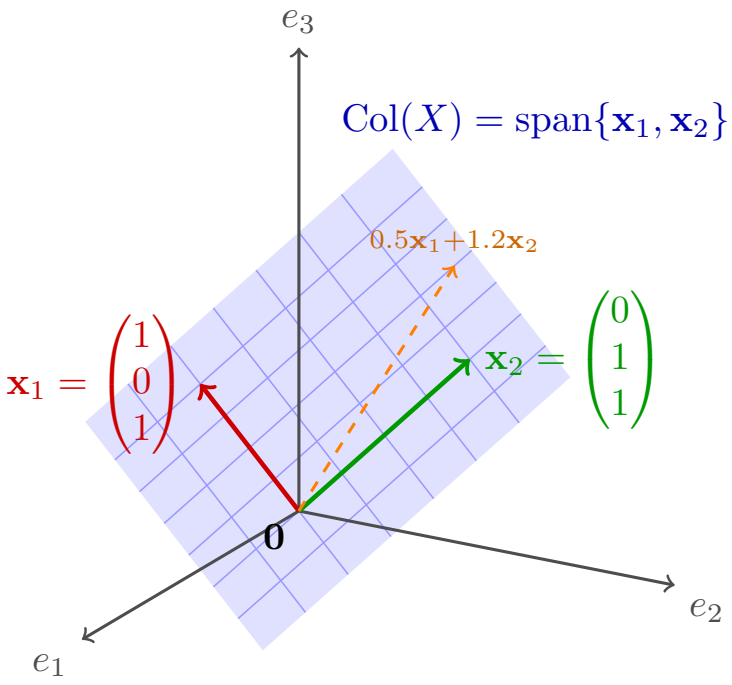

Figure 1: Geometric visualization of the column space $\mathrm{Col}(X)$ for a 3×2 data matrix. The vectors $\mathbf{x}_1 = (1, 0, 1)^\top$ (red) and $\mathbf{x}_2 = (0, 1, 1)^\top$ (green) are the columns of X and form a basis for the column space. The shaded blue region represents the plane spanned by these vectors, which constitutes all possible fitted values $\hat{\mathbf{y}} = X\boldsymbol{\beta}$. The dashed orange vector illustrates a sample linear combination $0.5\mathbf{x}_1 + 1.2\mathbf{x}_2$, demonstrating that any such combination lies within the plane.

Geometrically, this is a two-dimensional plane passing through the origin in $\mathbb{R}^3$. Since $\mathbf{x}_1$ and $\mathbf{x}_2$ are linearly independent, we have $\mathrm{rank}(X) = 2$, and thus $\dim(\mathrm{Col}(X)) = 2$.

Figure 1 provides a geometric interpretation of the column space for the design matrix X in Example 1. The two column vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ serve as basis vectors that span a two-dimensional plane passing through the origin in $\mathbb{R}^3$. Every point on this plane corresponds to a unique linear combination $\beta_1\mathbf{x}_1+\beta_2\mathbf{x}_2$ for some choice of coefficients $\beta_1, \beta_2 \in \mathbb{R}$. In the context of linear regression, this plane represents the set of all possible fitted values $\hat{\mathbf{y}} = X\boldsymbol{\beta}$ that can be produced by varying the parameter vector $\boldsymbol{\beta}$. Crucially, if the response vector $\mathbf{y}$ does not lie on this plane, no exact solution to $X\boldsymbol{\beta} = \mathbf{y}$ exists, and we must instead seek the best approximation—a problem naturally solved by orthogonal projection.

Example 1.2. Consider a matrix where the columns are linearly dependent:

$$X = \begin{pmatrix}1 & 2\\2 & 4\\3 & 6\end{pmatrix}, \quad \mathbf{x}_1 = \begin{pmatrix}1\\2\\3\end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix}2\\4\\6\end{pmatrix} = 2\mathbf{x}_1.$$

Although X has two columns, the column space is:

$$\mathrm{Col}(X) = \mathrm{span}\{\mathbf{x}_1, \mathbf{x}_2\} = \mathrm{span}\{\mathbf{x}_1\} = \left\{ t \begin{pmatrix}1\\2\\3\end{pmatrix} : t \in \mathbb{R} \right\}.$$

Geometrically, this is only a one-dimensional line through the origin in $\mathbb{R}^3$. Here, $\text{rank}(X) = 1$, illustrating that the dimension of the column space equals the rank of the matrix, not necessarily the number of columns. This situation, known as **multicollinearity**, causes problems in regression since $X^\top X$ becomes singular.

The remainder of this lecture is dedicated to formalizing this assertion, deriving its algebraic consequences, and proving its equivalence to the traditional analytical formulation of the problem.

2 The Algebraic Formulation

2.1 The Linear Model in Matrix Form

Let us begin with the familiar algebraic statement of a linear model for a single observation i:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{1}$$

Here, the β_j are the unknown model coefficients we wish to estimate, and ϵ_i is the irreducible error term for observation i, representing noise and unmodeled phenomena.

To express this system for all n observations simultaneously, and to elegantly handle the intercept term β_0, we augment our feature matrix $X \in \mathbb{R}^{n \times p}$ with a leading column of ones. Let us define this column as $\mathbf{x}_0 = [1, 1, \ldots, 1]^T \in \mathbb{R}^n$. Our new matrix, which we will continue to call X for simplicity, is now the $n \times (p+1)$ **design matrix**. Correspondingly, our vector of coefficients becomes $\beta \in \mathbb{R}^{p+1}$, with $\beta = [\beta_0, \beta_1, \ldots, \beta_p]^T$.

This augmentation allows us to write the entire system of n linear equations (1) in a single, compact matrix equation

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Or, more succinctly

$$\mathbf{y} = X\beta + \epsilon.$$

2.2 The Impossibility of an Exact Solution

In an ideal, noise-free world, we would seek a coefficient vector β that provides a perfect solution, such that $\mathbf{y} = X\beta$. This represents a system of n linear equations in $p+1$ unknown coefficients.

However, in virtually all practical applications of data analysis, the number of observations n is significantly larger than the number of features p (i.e., $n \gg p+1$). Such a system is said to be **overdetermined**. An exact solution to the equation $X\beta = \mathbf{y}$ exists if and only if the vector $\mathbf{y}$ is a linear combination of the columns of X; that is, if $\mathbf{y} \in \text{Col}(X)$.

The presence of the error vector ϵ makes this possibility vanishingly small. This vector ϵ encapsulates all sources of deviation from a perfect linear relationship: measurement noise, the influence of unobserved variables, and intrinsic randomness in the underlying process. Consequently, the observed target vector $\mathbf{y}$ will almost certainly *not* lie within the comparatively small subspace $\text{Col}(X)$ spanned by the feature vectors.

Geometrically, the vector $\mathbf{y}$ *sticks out* of the hyperplane defined by $\text{Col}(X)$. No matter what linear combination of the columns of X we choose (i.e., no matter what β we select), we cannot form

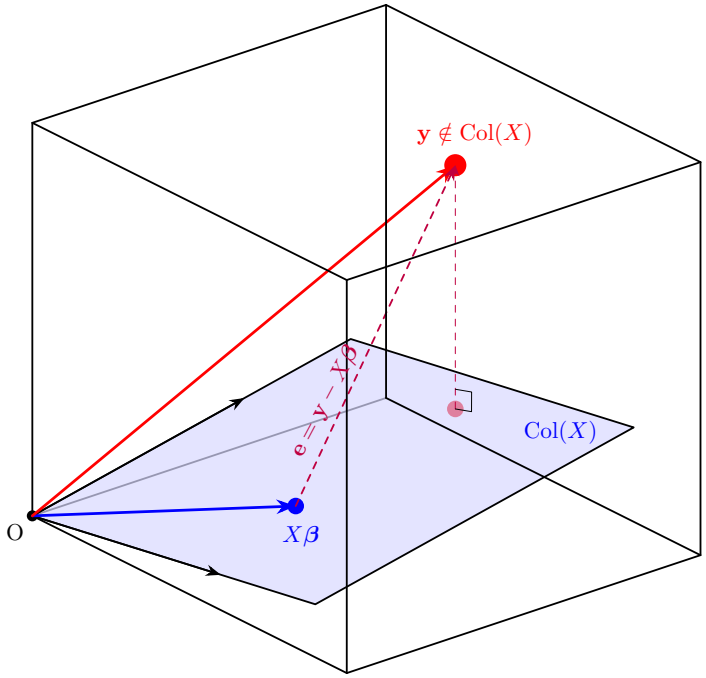

Figure 2: Geometric illustration of why an exact solution to $X\boldsymbol{\beta} = \mathbf{y}$ is impossible in overdetermined systems.

a vector $X\beta$ that perfectly equals $\mathbf{y}$. The equation $\mathbf{y} = X\beta + \epsilon$ can be rearranged to $\epsilon = \mathbf{y} - X\beta$. This provides a direct geometric meaning to the error term: the vector ϵ is the displacement vector that connects a point $X\beta$ *within* the column space to the observed data vector $\mathbf{y}$ which lies *outside* it. The impossibility of an exact solution is equivalent to the geometric statement that this displacement vector is non-zero for any choice of β. Our task, therefore, is not to find a β that makes this vector zero, but to find the specific $\hat{\beta}$ that makes this displacement vector as short as possible.

Figure 2 provides a geometric interpretation of why an exact solution to the linear system $X\boldsymbol{\beta} = \mathbf{y}$ is generally unattainable in regression problems. The column space $\mathrm{Col}(X)$, depicted as the shaded plane, represents the set of all possible fitted values achievable by any choice of coefficient vector $\boldsymbol{\beta}$. However, the observed response vector $\mathbf{y}$ (shown in purple) typically does not reside within this subspace; it sticks out of the plane due to noise, measurement error, and unmodeled phenomena. The displacement vector $\mathbf{e} = \mathbf{y} - X\boldsymbol{\beta}$ (in purple) connects any candidate fitted value in the column space to the true response outside it. Since $\mathbf{y} \notin \mathrm{Col}(X)$, this displacement can never be zero. The goal of least squares regression is therefore to find the coefficient vector $\hat{\boldsymbol{\beta}}$ that minimizes the length of this displacement—geometrically, this corresponds to dropping a perpendicular from $\mathbf{y}$ onto the plane.

3 The Orthogonal Projection Principle

Before presenting the geometric solution to the linear regression problem, we establish the foundational concepts from linear algebra that underpin this approach. These definitions formalize the notions of *distance* and *perpendicularity* that are central to our geometric interpretation.

3.1 Preliminaries

Definition 3.1 (Inner Product)**.** Let V be a vector space over $\mathbb{R}$. An **inner product** on V is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies the following axioms for all vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and all scalars $\alpha \in \mathbb{R}$:

1. *Symmetry*: $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$
2. *Linearity in the first argument*: $\langle \alpha\mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \alpha\langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$
3. *Positive definiteness*: $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$, with equality if and only if $\mathbf{u} = \mathbf{0}$

A vector space equipped with an inner product is called an **inner product space**.

Example 3.2. The canonical example relevant to our study is $\mathbb{R}^n$ equipped with the **Euclidean inner product** (or dot product):

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v} = \sum_{i=1}^{n} u_i v_i.$$

This is the inner product we employ throughout our treatment of linear regression in the observation space.

Example 3.3. Another important example arises in the study of function spaces. Consider the vector space $C([a, b])$ of continuous real-valued functions on the interval $[a, b]$. This space can be equipped with the L^2 inner product:

$$\langle f, g \rangle = \int_a^b f(x)g(x)\, dx.$$

Under this inner product, two functions are orthogonal if their pointwise product integrates to zero over the interval. For instance, the functions $f(x) = \sin(x)$ and $g(x) = \cos(x)$ are orthogonal on $[0, 2\pi]$ since

$$\langle f, g \rangle = \int_0^{2\pi} \sin(x)\cos(x)\, dx = \frac{1}{2}\sin^2(x)\Big|_0^{2\pi} = 0.$$

This inner product space provides the theoretical foundation for Fourier analysis and is essential in understanding kernel methods in machine learning, where data is implicitly mapped into infinite-dimensional function spaces.

Definition 3.4 (Induced Norm)**.** Let V be an inner product space. The norm induced by the inner product is the function $\| \cdot \| : V \to \mathbb{R}$ defined by

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}.$$

This norm measures the *length* or *magnitude* of a vector. In $\mathbb{R}^n$ with the Euclidean inner product, this yields the familiar Euclidean norm:

$$\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^\top \mathbf{v}} = \sqrt{\sum_{i=1}^{n} v_i^2}.$$

The **distance** between two vectors $\mathbf{u}$ and $\mathbf{v}$ is defined as $\|\mathbf{u} - \mathbf{v}\|$.

Example 3.5. Consider the vector $\mathbf{v} = (3, -4, 0)^\top \in \mathbb{R}^3$ with the Euclidean inner product. The induced norm is

$$\|\mathbf{v}\|_2 = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{3^2 + (-4)^2 + 0^2} = \sqrt{9 + 16 + 0} = \sqrt{25} = 5.$$

Now consider a second vector $\mathbf{u} = (1, 0, 2)^\top$. The distance between $\mathbf{u}$ and $\mathbf{v}$ is

$$\|\mathbf{u} - \mathbf{v}\| = \|(1 - 3, 0 - (-4), 2 - 0)^\top\| = \|(-2, 4, 2)^\top\| = \sqrt{4 + 16 + 4} = \sqrt{24} = 2\sqrt{6}.$$

In the context of linear regression, this distance formula is precisely what we seek to minimize: the Euclidean distance between the observed response vector $\mathbf{y}$ and the fitted vector $\hat{\mathbf{y}}$.

Example 3.6. In the function space $C([0, 1])$ equipped with the L^2 inner product, consider the function $f(x) = x$. Its induced norm is

$$\|f\| = \sqrt{\langle f, f \rangle} = \sqrt{\int_0^1 x^2 \, dx} = \sqrt{\frac{x^3}{3}\Big|_0^1} = \sqrt{\frac{1}{3}} = \frac{1}{\sqrt{3}}.$$

Similarly, for the constant function $g(x) = 1$, we have $\|g\| = \sqrt{\int_0^1 1 \, dx} = 1$. The distance between f and g is

$$\|f - g\| = \sqrt{\int_0^1 (x - 1)^2 \, dx} = \sqrt{\int_0^1 (x^2 - 2x + 1) \, dx} = \sqrt{\frac{1}{3} - 1 + 1} = \frac{1}{\sqrt{3}}.$$

This measures how *far apart* the two functions are in the L^2 sense—not pointwise, but in an averaged, integral sense over the entire domain.

Definition 3.7 (Orthogonality)**.** Two vectors $\mathbf{u}, \mathbf{v} \in V$ are said to be **orthogonal** (or perpendicular), denoted $\mathbf{u} \perp \mathbf{v}$, if their inner product vanishes:

$$\mathbf{u} \perp \mathbf{v} \iff \langle \mathbf{u}, \mathbf{v} \rangle = 0.$$

A vector $\mathbf{v}$ is **orthogonal to a subspace** $W \subseteq V$, denoted $\mathbf{v} \perp W$, if $\mathbf{v}$ is orthogonal to every vector in W:

$$\mathbf{v} \perp W \iff \langle \mathbf{v}, \mathbf{w} \rangle = 0 \quad \text{for all } \mathbf{w} \in W.$$

Example 3.8. In $\mathbb{R}^3$ with the Euclidean inner product, consider the vectors $\mathbf{u} = (1, 2, -1)^\top$ and $\mathbf{v} = (3, 0, 3)^\top$. We verify orthogonality by computing their inner product:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v} = (1)(3) + (2)(0) + (-1)(3) = 3 + 0 - 3 = 0.$$

Since $\langle \mathbf{u}, \mathbf{v} \rangle = 0$, we conclude that $\mathbf{u} \perp \mathbf{v}$. Geometrically, these vectors meet at a right angle in three-dimensional space.

Now consider the subspace $W = \operatorname{span}\{(1, 0, 0)^\top, (0, 1, 0)^\top\}$, which is the xy-plane. The vector $\mathbf{n} = (0, 0, 1)^\top$ is orthogonal to W because for any $\mathbf{w} = (a, b, 0)^\top \in W$:

$$\langle \mathbf{n}, \mathbf{w} \rangle = (0)(a) + (0)(b) + (1)(0) = 0.$$

The vector $\mathbf{n}$ is the normal vector to the plane W, illustrating that orthogonality to a subspace generalizes the familiar notion of perpendicularity to a plane.

Example 3.9. This example foreshadows a key result in linear regression. Consider a design matrix $X \in \mathbb{R}^{3\times 2}$ and a residual vector $\mathbf{e} \in \mathbb{R}^3$:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

The column space of X is $W = \mathrm{Col}(X) = \mathrm{span}\{\mathbf{x}_0, \mathbf{x}_1\}$ where $\mathbf{x}_0 = (1,1,1)^\top$ and $\mathbf{x}_1 = (1,2,3)^\top$. To verify that $\mathbf{e} \perp W$, it suffices to check orthogonality against the spanning vectors:

$$\langle \mathbf{x}_0, \mathbf{e} \rangle = (1)(1) + (1)(-2) + (1)(1) = 1 - 2 + 1 = 0,$$

$$\langle \mathbf{x}_1, \mathbf{e} \rangle = (1)(1) + (2)(-2) + (3)(1) = 1 - 4 + 3 = 0.$$

Since $\mathbf{e}$ is orthogonal to every basis vector of W, we have $\mathbf{e} \perp W$. These two conditions can be written compactly as $X^\top \mathbf{e} = \mathbf{0}$, which is precisely the orthogonality condition that defines the optimal residual in least squares regression.

Definition 3.10 (Orthogonal Complement)**.** Let W be a subspace of an inner product space V. The **orthogonal complement** of W, denoted $W^\perp$, is the set of all vectors in V that are orthogonal to every vector in W:

$$W^\perp = \{\mathbf{v} \in V : \langle \mathbf{v}, \mathbf{w} \rangle = 0 \text{ for all } \mathbf{w} \in W\}.$$

The orthogonal complement $W^\perp$ is itself a subspace of V.

Example 3.11. In $\mathbb{R}^3$, let $W = \mathrm{span}\{(1,0,1)^\top\}$ be the one-dimensional subspace (a line through the origin). To find $W^\perp$, we seek all vectors $\mathbf{v} = (v_1, v_2, v_3)^\top$ such that $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ for all $\mathbf{w} \in W$. Since W is spanned by a single vector, it suffices to require orthogonality to that generator:

$$\langle \mathbf{v}, (1,0,1)^\top \rangle = v_1 + v_3 = 0 \implies v_3 = -v_1.$$

Thus, $W^\perp = \{(v_1, v_2, -v_1)^\top : v_1, v_2 \in \mathbb{R}\} = \mathrm{span}\{(1,0,-1)^\top, (0,1,0)^\top\}$. Geometrically, W is a line and $W^\perp$ is the plane passing through the origin that is perpendicular to this line. Note that $\dim(W) + \dim(W^\perp) = 1 + 2 = 3 = \dim(\mathbb{R}^3)$, illustrating the general dimension formula.

Definition 3.12 (Left Null Space)**.** Let $X \in \mathbb{R}^{n\times p}$ be a matrix. The **left null space** of X, denoted $\mathrm{Null}(X^\top)$, is the set of all vectors in $\mathbb{R}^n$ that are mapped to zero by $X^\top$:

$$\mathrm{Null}(X^\top) = \{\mathbf{v} \in \mathbb{R}^n : X^\top \mathbf{v} = \mathbf{0}\}.$$

Example 3.13. Consider the matrix

$$X = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{pmatrix} \in \mathbb{R}^{3\times 2}.$$

Since the second column is twice the first, we have $\mathrm{rank}(X) = 1$, and thus $\dim(\mathrm{Null}(X^\top)) = 3 - 1 = 2$. To find the left null space, we solve $X^\top \mathbf{v} = \mathbf{0}$:

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies v_1 + 2v_2 + 3v_3 = 0.$$

This single equation in three unknowns yields a two-dimensional solution space:

$$\text{Null}(X^\top) = \text{span}\left\{ \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

Geometrically, since $\text{Col}(X)$ is a line through the origin in $\mathbb{R}^3$, its orthogonal complement $\text{Null}(X^\top)$ is the plane perpendicular to that line.

Example 3.14. Consider the design matrix from our earlier example:

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \text{with } W = \text{Col}(X) = \text{span}\left\{ \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right\} \subseteq \mathbb{R}^3.$$

To find $W^\perp$, we need all vectors $\mathbf{v} = (v_1, v_2, v_3)^\top$ orthogonal to both columns of X:

$$v_1 + v_3 = 0 \quad \text{and} \quad v_2 + v_3 = 0.$$

Solving this system yields $v_1 = v_2 = -v_3$, so $W^\perp = \text{span}\{(1, 1, -1)^\top\}$. This one-dimensional subspace is perpendicular to the column space plane.

Theorem 3.15. *Let $X \in \mathbb{R}^{n \times p}$ be a matrix and let $\mathbf{v} \in \mathbb{R}^n$. The following statements are equivalent:*

1. $\mathbf{v} \in \text{Null}(X^\top)$.
2. $\mathbf{v}^\top X = \mathbf{0}^\top$.

Proof. We establish the equivalence by showing that each statement implies the other.

Suppose first that $\mathbf{v} \in \text{Null}(X^\top)$, i.e., $X^\top \mathbf{v} = \mathbf{0}$. Taking the transpose of both sides, we obtain

$$(X^\top \mathbf{v})^\top = \mathbf{0}^\top.$$

The left-hand side simplifies as $(X^\top \mathbf{v})^\top = \mathbf{v}^\top (X^\top)^\top = \mathbf{v}^\top X$. Therefore, $\mathbf{v}^\top X = \mathbf{0}^\top$.

Conversely, suppose that $\mathbf{v}^\top X = \mathbf{0}^\top$. Taking the transpose of both sides yields

$$(\mathbf{v}^\top X)^\top = (\mathbf{0}^\top)^\top.$$

The left-hand side simplifies as $(\mathbf{v}^\top X)^\top = X^\top (\mathbf{v}^\top)^\top = X^\top \mathbf{v}$, while the right-hand side is simply $\mathbf{0}$. Therefore, $X^\top \mathbf{v} = \mathbf{0}$. □

This equivalence justifies the terminology *left null space*: a vector $\mathbf{v}$ belongs to this subspace if and only if it annihilates X when multiplying from the left (i.e., $\mathbf{v}^\top X = \mathbf{0}^\top$). In contrast, the ordinary null space $\text{Null}(X)$ consists of vectors that annihilate X when multiplying from the right (i.e., $X\mathbf{w} = \mathbf{0}$).

Theorem 3.16. *Let $X \in \mathbb{R}^{n \times p}$ be a matrix. The left null space of X is equal to the orthogonal complement of the column space of X:*

$$\text{Null}(X^\top) = \text{Col}(X)^\perp.$$

Proof. To show that $\text{Null}(X^\top) \subseteq \text{Col}(X)^\perp$, let $\mathbf{v}$ be an arbitrary vector in $\text{Null}(X^\top)$. By definition, this means $X^\top \mathbf{v} = \mathbf{0}$. We must show that $\mathbf{v}$ is orthogonal to every vector in $\text{Col}(X)$. Let $\mathbf{w} \in \text{Col}(X)$ be arbitrary. By definition of the column space, there exists some $\boldsymbol{\alpha} \in \mathbb{R}^p$ such that $\mathbf{w} = X\boldsymbol{\alpha}$. Computing the inner product of $\mathbf{v}$ with $\mathbf{w}$, we obtain

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^\top \mathbf{w} = \mathbf{v}^\top (X\boldsymbol{\alpha}) = (\mathbf{v}^\top X)\boldsymbol{\alpha} = (X^\top \mathbf{v})^\top \boldsymbol{\alpha} = \mathbf{0}^\top \boldsymbol{\alpha} = 0.$$

Since $\mathbf{w}$ was an arbitrary element of $\text{Col}(X)$, we conclude that $\mathbf{v} \perp \text{Col}(X)$, and hence $\mathbf{v} \in \text{Col}(X)^\perp$.

To show the reverse inclusion $\text{Col}(X)^\perp \subseteq \text{Null}(X^\top)$, let $\mathbf{v}$ be an arbitrary vector in $\text{Col}(X)^\perp$. By definition, $\mathbf{v}$ is orthogonal to every vector in $\text{Col}(X)$. In particular, $\mathbf{v}$ must be orthogonal to each column of X. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ denote the columns of X. Then for each $j \in \{1, 2, \ldots, p\}$, we have $\mathbf{x}_j \in \text{Col}(X)$, and therefore

$$\langle \mathbf{v}, \mathbf{x}_j \rangle = \mathbf{v}^\top \mathbf{x}_j = 0.$$

The product $X^\top \mathbf{v}$ is a vector in $\mathbb{R}^p$ whose j-th component is precisely $\mathbf{x}_j^\top \mathbf{v} = \mathbf{v}^\top \mathbf{x}_j$. Since each of these components vanishes, we have

$$X^\top \mathbf{v} = \begin{pmatrix} \mathbf{x}_1^\top \mathbf{v} \\ \mathbf{x}_2^\top \mathbf{v} \\ \vdots \\ \mathbf{x}_p^\top \mathbf{v} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}.$$

This shows that $\mathbf{v} \in \text{Null}(X^\top)$, establishing the second inclusion.

Having demonstrated both $\text{Null}(X^\top) \subseteq \text{Col}(X)^\perp$ and $\text{Col}(X)^\perp \subseteq \text{Null}(X^\top)$, we conclude that $\text{Null}(X^\top) = \text{Col}(X)^\perp$. □

Definition 3.17 (Direct Sum)**.** Let U and W be subspaces of a vector space V. The vector space V is said to be the **direct sum** of U and W, denoted $V = U \oplus W$, if the following two conditions are satisfied:

1. *Spanning property*: Every vector $\mathbf{v} \in V$ can be written as $\mathbf{v} = \mathbf{u} + \mathbf{w}$ for some $\mathbf{u} \in U$ and $\mathbf{w} \in W$.

2. *Trivial intersection*: The subspaces intersect only at the origin, i.e., $U \cap W = \{\mathbf{0}\}$.

When both conditions hold, every vector $\mathbf{v} \in V$ has a *unique* representation as $\mathbf{v} = \mathbf{u} + \mathbf{w}$ with $\mathbf{u} \in U$ and $\mathbf{w} \in W$. Furthermore, if $V = U \oplus W$, then $\dim(V) = \dim(U) + \dim(W)$.

Example 3.18. Consider $V = \mathbb{R}^2$ and define the subspaces

$$U = \text{span}\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} = \{(x, 0)^\top : x \in \mathbb{R}\}, \quad W = \text{span}\left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} = \{(0, y)^\top : y \in \mathbb{R}\}.$$

Geometrically, U is the x-axis and W is the y-axis. We verify the two conditions for a direct sum:

1. *Spanning*: Any vector $(a, b)^\top \in \mathbb{R}^2$ can be written as $(a, b)^\top = (a, 0)^\top + (0, b)^\top$ with $(a, 0)^\top \in U$ and $(0, b)^\top \in W$.

2. *Trivial intersection*: If $\mathbf{v} \in U \cap W$, then $\mathbf{v} = (x, 0)^\top$ for some x and $\mathbf{v} = (0, y)^\top$ for some y. This forces $x = 0$ and $y = 0$, so $\mathbf{v} = \mathbf{0}$.

Therefore, $\mathbb{R}^2 = U \oplus W$. Note that $\dim(\mathbb{R}^2) = 2 = 1 + 1 = \dim(U) + \dim(W)$, confirming the dimension formula. This decomposition corresponds to expressing any point in the plane via its Cartesian coordinates.

Example 3.19. Consider $V = \mathbb{R}^3$ and define the subspaces

$$U = \text{span}\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right\}, \quad W = \text{span}\left\{ \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \right\}.$$

Here U is a plane through the origin and W is a line through the origin. We have $\dim(U) = 2$ and $\dim(W) = 1$.

To verify that $\mathbb{R}^3 = U \oplus W$, we first check the trivial intersection property. Suppose $\mathbf{v} \in U \cap W$. Then $\mathbf{v} = \alpha(1,1,0)^\top + \beta(0,1,1)^\top = \gamma(1,-1,1)^\top$ for some scalars α, β, γ. This yields the system:

$$\alpha = \gamma, \quad \alpha + \beta = -\gamma, \quad \beta = \gamma.$$

From the first and third equations, $\alpha = \beta = \gamma$. Substituting into the second equation: $\gamma + \gamma = -\gamma$, which gives $3\gamma = 0$, so $\gamma = 0$. Hence $\mathbf{v} = \mathbf{0}$, confirming $U \cap W = \{\mathbf{0}\}$.

For the spanning property, since $U \cap W = \{\mathbf{0}\}$ and $\dim(U) + \dim(W) = 2 + 1 = 3 = \dim(\mathbb{R}^3)$, linear algebra guarantees that every vector in $\mathbb{R}^3$ can be expressed as a sum of vectors from U and W. For instance, the standard basis vector $\mathbf{e}_1 = (1,0,0)^\top$ can be decomposed by solving

$$\alpha \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \beta \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + \gamma \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

yielding $\alpha = \frac{1}{3}$, $\beta = -\frac{1}{3}$, $\gamma = \frac{2}{3}$. Thus $\mathbf{e}_1 = \underbrace{\frac{1}{3}(1,1,0)^\top - \frac{1}{3}(0,1,1)^\top}_{\in U} + \underbrace{\frac{2}{3}(1,-1,1)^\top}_{\in W}$.

Therefore, $\mathbb{R}^3 = U \oplus W$. Note that U and W are *not* orthogonal complements of each other in this example; the direct sum structure does not require orthogonality. However, when $W = U^\perp$, the decomposition $V = U \oplus U^\perp$ is called an *orthogonal direct sum*.

The following theorem establishes a fundamental decomposition property that is essential to understanding projections.

Theorem 3.20 (Orthogonal Decomposition)**.** *Let W be a finite-dimensional subspace of an inner product space V. It follows that*

$$V = W \oplus W^\perp.$$

Proof. To establish that $V = W \oplus W^\perp$, we must verify the two defining conditions of a direct sum: the spanning property and the trivial intersection property.

We begin by showing that the only vector belonging to both W and $W^\perp$ is the zero vector. Suppose $\mathbf{z} \in W \cap W^\perp$. Since $\mathbf{z} \in W^\perp$, by definition $\mathbf{z}$ is orthogonal to every vector in W. In particular, since $\mathbf{z} \in W$, the vector $\mathbf{z}$ must be orthogonal to itself:

$$\langle \mathbf{z}, \mathbf{z} \rangle = 0.$$

By the positive definiteness axiom of the inner product, $\langle \mathbf{z}, \mathbf{z} \rangle = 0$ implies $\mathbf{z} = \mathbf{0}$. Therefore, $W \cap W^\perp = \{\mathbf{0}\}$, establishing the trivial intersection property.

We now show that every vector in V can be expressed as the sum of a vector in W and a vector in $W^\perp$. Since W is finite-dimensional, it admits an orthonormal basis. Let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ be such a basis, where $k = \dim(W)$. The existence of an orthonormal basis is guaranteed by the Gram-Schmidt process applied to any basis of W.

Let $\mathbf{v} \in V$ be an arbitrary vector. We construct the orthogonal projection of $\mathbf{v}$ onto W by defining

$$\mathbf{w} = \sum_{j=1}^{k} \langle \mathbf{v}, \mathbf{u}_j \rangle \, \mathbf{u}_j.$$

Since $\mathbf{w}$ is a linear combination of the basis vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$, and since W is closed under linear combinations, we have $\mathbf{w} \in W$.

We now define $\mathbf{w}^\perp = \mathbf{v} - \mathbf{w}$ and claim that $\mathbf{w}^\perp \in W^\perp$. To verify this claim, we must show that $\langle \mathbf{w}^\perp, \mathbf{z} \rangle = 0$ for all $\mathbf{z} \in W$. Since the set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ spans W, it suffices to verify orthogonality against each basis vector. Fix an arbitrary index $i \in \{1, 2, \ldots, k\}$ and compute:

$$\begin{aligned}
\langle \mathbf{w}^\perp, \mathbf{u}_i \rangle &= \langle \mathbf{v} - \mathbf{w}, \mathbf{u}_i \rangle \\
&= \langle \mathbf{v}, \mathbf{u}_i \rangle - \langle \mathbf{w}, \mathbf{u}_i \rangle \\
&= \langle \mathbf{v}, \mathbf{u}_i \rangle - \left\langle \sum_{j=1}^{k} \langle \mathbf{v}, \mathbf{u}_j \rangle \, \mathbf{u}_j, \; \mathbf{u}_i \right\rangle \\
&= \langle \mathbf{v}, \mathbf{u}_i \rangle - \sum_{j=1}^{k} \langle \mathbf{v}, \mathbf{u}_j \rangle \langle \mathbf{u}_j, \mathbf{u}_i \rangle.
\end{aligned}$$

The final step follows from the linearity of the inner product in the first argument. By orthonormality of the basis, $\langle \mathbf{u}_j, \mathbf{u}_i \rangle = \delta_{ij}$, the Kronecker delta, which equals unity when $j = i$ and vanishes otherwise. Consequently, the summation reduces to a single term:

$$\langle \mathbf{w}^\perp, \mathbf{u}_i \rangle = \langle \mathbf{v}, \mathbf{u}_i \rangle - \langle \mathbf{v}, \mathbf{u}_i \rangle \cdot 1 = 0.$$

Since this holds for every $i \in \{1, \ldots, k\}$, and since any $\mathbf{z} \in W$ can be written as $\mathbf{z} = \sum_{i=1}^{k} c_i \mathbf{u}_i$ for some scalars c_i, the linearity of the inner product yields

$$\langle \mathbf{w}^\perp, \mathbf{z} \rangle = \sum_{i=1}^{k} c_i \langle \mathbf{w}^\perp, \mathbf{u}_i \rangle = \sum_{i=1}^{k} c_i \cdot 0 = 0.$$

Therefore, $\mathbf{w}^\perp \perp W$, which means $\mathbf{w}^\perp \in W^\perp$.

We have thus expressed the arbitrary vector $\mathbf{v} \in V$ as

$$\mathbf{v} = \mathbf{w} + \mathbf{w}^\perp,$$

where $\mathbf{w} \in W$ and $\mathbf{w}^\perp \in W^\perp$. This establishes the spanning property $V = W + W^\perp$.

□

Definition 3.21 (Orthogonal Projection)**.** Let W be a finite-dimensional subspace of an inner product space V, and let $\mathbf{y} \in V$. The **orthogonal projection** of $\mathbf{y}$ onto W, denoted $\mathrm{proj}_W(\mathbf{y})$, is the unique vector in W such that

$$\mathbf{y} - \mathrm{proj}_W(\mathbf{y}) \in W^\perp,$$

Equivalently, $\mathbf{y} - \mathrm{proj}_W(\mathbf{y})$ is orthogonal to every vector in W. By the Orthogonal Decomposition Theorem, if $\mathbf{y} = \mathbf{w} + \mathbf{w}^\perp$ is the unique decomposition with $\mathbf{w} \in W$ and $\mathbf{w}^\perp \in W^\perp$, then $\mathrm{proj}_W(\mathbf{y}) = \mathbf{w}$.

The following classical result connects orthogonality to the Euclidean notion of distance and underpins the geometric interpretation of projections.

Theorem 3.22 (Pythagorean Theorem)**.** *Let V be an inner product space. If $\mathbf{u}, \mathbf{v} \in V$ are orthogonal, i.e., $\langle \mathbf{u}, \mathbf{v} \rangle = 0$, then*

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2.$$

Proof. By expanding the squared norm using the inner product:

$$\|\mathbf{u} + \mathbf{v}\|^2 = \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + 2\langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{u}\|^2 + 0 + \|\mathbf{v}\|^2.$$

□

When the underlying vector space is finite-dimensional, any linear transformation can be represented by a matrix. It is natural to ask: what properties must a matrix possess in order to represent an orthogonal projection onto some subspace? The answer lies in two elegant algebraic conditions—symmetry and idempotence—that together fully characterize orthogonal projection matrices.

Definition 3.23 (Orthogonal Projection Matrix)**.** Let $V = \mathbb{R}^n$ be equipped with the Euclidean inner product. A matrix $P \in \mathbb{R}^{n \times n}$ is called an **orthogonal projection matrix** if it satisfies the following two properties:

1. *Symmetry*: $P^\top = P$.
2. *Idempotence*: $P^2 = P$.

The symmetry condition ensures that the projection respects the inner product structure: for any vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we have $\langle P\mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, P\mathbf{v} \rangle$. The idempotence condition captures the intuitive notion that projecting twice yields the same result as projecting once—if a vector already lies in the target subspace, the projection leaves it unchanged. Together, these properties guarantee that P decomposes $\mathbb{R}^n$ into orthogonal components: for any $\mathbf{v} \in \mathbb{R}^n$, the decomposition $\mathbf{v} = P\mathbf{v} + (I - P)\mathbf{v}$ satisfies $P\mathbf{v} \in \mathrm{Col}(P)$ and $(I - P)\mathbf{v} \in \mathrm{Col}(P)^\perp$, with the two components orthogonal to each other.

Example 3.24. Consider the projection onto the one-dimensional subspace $W = \mathrm{span}\{(1, 1)^\top\}$ in $\mathbb{R}^2$. This subspace is the line through the origin with slope 1. To construct the projection matrix, we use the formula $P = \mathbf{u}\mathbf{u}^\top$ where $\mathbf{u}$ is a unit vector spanning W. Normalizing the spanning vector, we obtain $\mathbf{u} = \frac{1}{\sqrt{2}}(1, 1)^\top$, and thus

$$P = \mathbf{u}\mathbf{u}^\top = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

We verify the two defining properties. For symmetry, $P^\top = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = P$. For idempotence,

$$P^2 = \frac{1}{4}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = P.$$

To illustrate the projection in action, consider $\mathbf{v} = (3, 1)^\top$. The projection onto W is

$$P\mathbf{v} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 4 \\ 4 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}.$$

The residual is $\mathbf{v} - P\mathbf{v} = (3, 1)^\top - (2, 2)^\top = (1, -1)^\top$. One can verify that this residual is orthogonal to the subspace: $\langle (1, -1)^\top, (1, 1)^\top \rangle = 1 - 1 = 0$.

Example 3.25. Consider the projection onto the two-dimensional subspace

$$W = \text{span}\{(1,0,1)^\top, (0,1,1)^\top\} \in \mathbb{R}^3,$$

which corresponds to the column space from our earlier regression example. Let X be the matrix with these vectors as columns:

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

The projection matrix onto $\text{Col}(X)$ is given by $P = X(X^\top X)^{-1}X^\top$. We compute each component:

$$X^\top X = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad (X^\top X)^{-1} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

Thus, the projection matrix is

$$P = X(X^\top X)^{-1}X^\top = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \cdot \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

The matrix P is visibly symmetric. To verify idempotence, one can compute P^2 directly and confirm that $P^2 = P$. The complementary projection matrix $M = I_3 - P$ projects onto $W^\perp = \text{Null}(X^\top)$:

$$M = I_3 - P = \frac{1}{3}\begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}.$$

Note that each row of M is proportional to $(1,1,-1)$, confirming that

$$\text{Col}(M) = \text{span}\{(1,1,-1)^\top\} = W^\perp,$$

consistent with our earlier computation of the orthogonal complement.

Theorem 3.26. *Let $P \in \mathbb{R}^{n\times n}$ be a symmetric and idempotent matrix. Then for every $\mathbf{v} \in \mathbb{R}^n$, the vector $P\mathbf{v}$ is the orthogonal projection of $\mathbf{v}$ onto the column space of P. That is,*

$$P\mathbf{v} = \text{proj}_{\text{Col}(P)}(\mathbf{v}).$$

Proof. To establish that $P\mathbf{v}$ is the orthogonal projection of $\mathbf{v}$ onto $W = \text{Col}(P)$, we must verify two conditions: first, that $P\mathbf{v} \in W$, and second, that the residual $\mathbf{v} - P\mathbf{v}$ is orthogonal to every vector in W.

We begin by showing that $P\mathbf{v} \in \text{Col}(P)$ for any $\mathbf{v} \in \mathbb{R}^n$. By definition, the column space of P consists of all vectors of the form $P\mathbf{u}$ for some $\mathbf{u} \in \mathbb{R}^n$. Taking $\mathbf{u} = \mathbf{v}$, we see immediately that $P\mathbf{v} \in \text{Col}(P)$.

We now establish that $\mathbf{v} - P\mathbf{v}$ is orthogonal to the entire column space of P. Let $\mathbf{w} \in \text{Col}(P)$ be arbitrary. By definition of the column space, there exists some $\mathbf{u} \in \mathbb{R}^n$ such that $\mathbf{w} = P\mathbf{u}$. We compute the inner product of $\mathbf{v} - P\mathbf{v}$ with $\mathbf{w}$:

$$\langle \mathbf{v} - P\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v} - P\mathbf{v}, P\mathbf{u} \rangle.$$

Using the property that $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \mathbf{b}$ for the Euclidean inner product, we write

$$\langle \mathbf{v} - P\mathbf{v}, P\mathbf{u} \rangle = (\mathbf{v} - P\mathbf{v})^\top (P\mathbf{u}) = \mathbf{v}^\top P\mathbf{u} - (P\mathbf{v})^\top (P\mathbf{u}).$$

For the first term, we use the symmetry of P (i.e., $P^\top = P$) to obtain

$$\mathbf{v}^\top P\mathbf{u} = \mathbf{v}^\top P^\top \mathbf{u} = (P\mathbf{v})^\top \mathbf{u}.$$

For the second term, we apply symmetry followed by idempotence (i.e., $P^2 = P$):

$$(P\mathbf{v})^\top (P\mathbf{u}) = \mathbf{v}^\top P^\top P\mathbf{u} = \mathbf{v}^\top PP\mathbf{u} = \mathbf{v}^\top P^2\mathbf{u} = \mathbf{v}^\top P\mathbf{u} = (P\mathbf{v})^\top \mathbf{u}.$$

Substituting these expressions back, we find

$$\langle \mathbf{v} - P\mathbf{v}, P\mathbf{u} \rangle = (P\mathbf{v})^\top \mathbf{u} - (P\mathbf{v})^\top \mathbf{u} = 0.$$

Since $\mathbf{w} = P\mathbf{u}$ was an arbitrary element of $\mathrm{Col}(P)$, we have shown that $\mathbf{v} - P\mathbf{v}$ is orthogonal to every vector in $\mathrm{Col}(P)$, which means $\mathbf{v} - P\mathbf{v} \in \mathrm{Col}(P)^\perp$.

We have now verified both defining conditions of orthogonal projection: $P\mathbf{v} \in \mathrm{Col}(P)$ and $\mathbf{v} - P\mathbf{v} \in \mathrm{Col}(P)^\perp$. By the uniqueness assertion of the Orthogonal Decomposition Theorem, there is exactly one vector in $\mathrm{Col}(P)$ with this property, namely the orthogonal projection of $\mathbf{v}$ onto $\mathrm{Col}(P)$. Therefore, $P\mathbf{v} = \mathrm{proj}_{\mathrm{Col}(P)}(\mathbf{v})$. □

With these foundational concepts in place, we are now prepared to state and prove the central theorem that justifies our geometric approach to linear regression.

3.2 The Best Approximation Theorem

The entire geometric approach to linear regression rests upon a cornerstone theorem of linear algebra. This theorem formally establishes that the orthogonal projection is the solution to our problem of finding the closest vector.

Theorem 3.27 (Best Approximation Theorem)**.** *Let W be a subspace of an inner product space V, and let $\mathbf{y} \in V$. The orthogonal projection of $\mathbf{y}$ onto W is the unique vector in W that is closest to $\mathbf{y}$. That is, for any vector $\mathbf{w} \in W$ such that $\mathbf{w} \neq proj_W(\mathbf{y})$, we have*

$$\|\mathbf{y} - \mathrm{proj}_W(\mathbf{y})\| < \|\mathbf{y} - \mathbf{w}\|.$$

Proof. Let $\hat{\mathbf{y}} = \mathrm{proj}_W(\mathbf{y})$ denote the orthogonal projection of $\mathbf{y}$ onto W. By the definition of orthogonal projection, the residual vector $\mathbf{y} - \hat{\mathbf{y}}$ lies in the orthogonal complement $W^\perp$, meaning $\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{z} \rangle = 0$ for all $\mathbf{z} \in W$.

Now let $\mathbf{w} \in W$ be an arbitrary vector. We seek to compare the distance $\|\mathbf{y} - \mathbf{w}\|$ with the distance $\|\mathbf{y} - \hat{\mathbf{y}}\|$. To facilitate this comparison, we decompose the difference $\mathbf{y} - \mathbf{w}$ by introducing the projection $\hat{\mathbf{y}}$ as an intermediate point, i.e.,

$$\mathbf{y} - \mathbf{w} = (\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \mathbf{w}).$$

Observe that the vector $\hat{\mathbf{y}} - \mathbf{w}$ belongs to W, since both $\hat{\mathbf{y}}$ and $\mathbf{w}$ are elements of W. Since $\mathbf{y} - \hat{\mathbf{y}} \in W^\perp$ and $\hat{\mathbf{y}} - \mathbf{w} \in W$, these two vectors are orthogonal to each other, that is, $\langle \mathbf{y} - \hat{\mathbf{y}}, \hat{\mathbf{y}} - \mathbf{w} \rangle = 0$.

The orthogonality of these two vectors permits the application of the Pythagorean theorem. Computing the squared norm of $\mathbf{y} - \mathbf{w}$, we obtain

$$\|\mathbf{y} - \mathbf{w}\|^2 = \|(\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \mathbf{w})\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \mathbf{w}\|^2.$$

Since norms are non-negative, the term $\|\hat{\mathbf{y}} - \mathbf{w}\|^2 \geq 0$, and consequently

$$\|\mathbf{y} - \mathbf{w}\|^2 \geq \|\mathbf{y} - \hat{\mathbf{y}}\|^2.$$

Taking square roots, which preserves the inequality for non-negative quantities, yields

$$\|\mathbf{y} - \mathbf{w}\| \geq \|\mathbf{y} - \hat{\mathbf{y}}\|.$$

It remains to establish uniqueness by showing that equality holds only when $\mathbf{w} = \hat{\mathbf{y}}$. The inequality $\|\mathbf{y} - \mathbf{w}\|^2 \geq \|\mathbf{y} - \hat{\mathbf{y}}\|^2$ becomes an equality if and only if $\|\hat{\mathbf{y}} - \mathbf{w}\|^2 = 0$. By positive definiteness of the norm, this occurs precisely when $\hat{\mathbf{y}} - \mathbf{w} = \mathbf{0}$, or equivalently, $\mathbf{w} = \hat{\mathbf{y}}$. Therefore, for any $\mathbf{w} \in W$ with $\mathbf{w} \neq \hat{\mathbf{y}}$, the inequality is strict, i.e.,

$$\|\mathbf{y} - \hat{\mathbf{y}}\| < \|\mathbf{y} - \mathbf{w}\|.$$

□

3.3 The Orthogonality Condition

Applying Theorem 3.27 to our regression problem, where our vector space is $V = \mathbb{R}^n$ equipped with the Euclidean inner product and our subspace is $W = \text{Col}(X)$, we arrive at a profound conclusion. The best possible linear prediction, denoted $\hat{\mathbf{y}} = \text{proj}_{\text{Col}(X)}(\mathbf{y})$, is the orthogonal projection of the observed outcome vector $\mathbf{y}$ onto the column space of X. Since $\hat{\mathbf{y}} \in \text{Col}(X)$, there exists a unique coefficient vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$ such that $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$.

The defining characteristic of this projection is that the *residual vector*,

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\hat{\boldsymbol{\beta}},$$

must be orthogonal to the entire subspace $\text{Col}(X)$. By Theorem 3.20, we have $\mathbf{e} \in \text{Col}(X)^{\perp}$. Invoking the relationship established earlier between the left null space and the orthogonal complement of the column space, we additionally have

$$\mathbf{e} \in \text{Col}(X)^{\perp} = \text{Null}(X^{\top}).$$

A vector is orthogonal to a subspace if and only if it is orthogonal to every vector in a spanning set for that subspace. A natural spanning set for $\text{Col}(X)$ is the set of its column vectors, $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p\}$, where $\mathbf{x}_0 = \mathbf{1}_n$ is the intercept column. This observation yields a set of $p + 1$ specific orthogonality conditions, expressed using the Euclidean inner product:

$$\begin{aligned}
\langle \mathbf{x}_0, \mathbf{e} \rangle = 0 &\implies \mathbf{x}_0^{\top}\mathbf{e} = 0, \\
\langle \mathbf{x}_1, \mathbf{e} \rangle = 0 &\implies \mathbf{x}_1^{\top}\mathbf{e} = 0, \\
&\vdots \\
\langle \mathbf{x}_p, \mathbf{e} \rangle = 0 &\implies \mathbf{x}_p^{\top}\mathbf{e} = 0.
\end{aligned}$$

Each of these equations states that the residual vector is orthogonal to the corresponding feature vector. In statistical terms, this orthogonality implies that the residuals are uncorrelated with each predictor variable. The first condition, $\mathbf{x}_0^{\top}\mathbf{e} = \mathbf{1}_n^{\top}\mathbf{e} = \sum_{i=1}^{n} e_i = 0$, has a particularly intuitive interpretation, i.e., the residuals must sum to zero, ensuring that the model neither systematically overpredicts nor underpredicts on average.

3.4 The Master Equation of Orthogonality

The $p+1$ individual scalar equations derived above can be consolidated into a single, elegant matrix equation. Recall that the transpose $X^\top$ has the column vectors $\mathbf{x}_j^\top$ as its rows:

$$X^\top = \begin{pmatrix} — & \mathbf{x}_0^\top & — \\ — & \mathbf{x}_1^\top & — \\ & \vdots & \\ — & \mathbf{x}_p^\top & — \end{pmatrix} \in \mathbb{R}^{(p+1)\times n}.$$

Multiplying $X^\top$ by the residual vector $\mathbf{e}$ yields a vector whose j-th component is precisely $\mathbf{x}_j^\top \mathbf{e}$:

$$X^\top \mathbf{e} = \begin{pmatrix} \mathbf{x}_0^\top \mathbf{e} \\ \mathbf{x}_1^\top \mathbf{e} \\ \vdots \\ \mathbf{x}_p^\top \mathbf{e} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}. \tag{2}$$

This equation is not merely a mathematical convenience; it carries a deep interpretation regarding information extraction. Each column $\mathbf{x}_j$ represents the information contained in a single feature across all observations. The inner product $\mathbf{x}_j^\top \mathbf{e}$ measures the linear association between that feature and the model's errors. The condition $X^\top \mathbf{e} = \mathbf{0}$ asserts that this association must vanish for *all* features simultaneously. In other words, the fitted values $\hat{\mathbf{y}}$ have extracted all possible linear information from the features in X to explain the target $\mathbf{y}$. The remaining error $\mathbf{e}$ is, by construction, linearly unpredictable from X. This provides a rigorous justification for calling $\hat{\mathbf{y}}$ the *best* linear fit, i.e., no further linear improvement is possible using the available features.

The master equation (2) is therefore the most compact statement of the geometric condition that the residual vector must be perpendicular to the subspace spanned by the predictors.

4 Derivation of the Estimator

We now possess all the geometric tools required to derive the algebraic form of the linear regression estimator. The derivation proceeds directly from the master orthogonality equation (2).

4.1 From Orthogonality to the Normal Equations

We know that our solution vector $\hat{\mathbf{y}}$ must lie in the column space of X. By the definition of the column space, there exists a coefficient vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$ such that

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} \in \mathrm{Col}(X).$$

This $\hat{\boldsymbol{\beta}}$ is precisely the vector of optimal coefficients we seek. Recalling that the residual vector is defined as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, we substitute the expression for $\hat{\mathbf{y}}$ into our orthogonality condition (2) to obtain

$$X^\top(\mathbf{y} - X\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Distributing the matrix $X^\top$ across the terms in the parentheses yields

$$X^\top \mathbf{y} - X^\top X\hat{\boldsymbol{\beta}} = \mathbf{0}.$$

Rearranging this equation by moving the second term to the right-hand side gives the celebrated **normal equation**

$$(X^\top X)\hat{\boldsymbol{\beta}} = X^\top \mathbf{y}. \tag{3}$$

The name *normal equations* is now revealed to be a direct consequence of our geometric derivation; it refers to the fact that the residual vector is required to be *normal* (i.e., perpendicular) to the column space of X. This terminology has nothing to do with the Gaussian (normal) distribution, though the two concepts are often conflated in elementary treatments.

4.2 The Gram Matrix

The matrix $G = X^\top X \in \mathbb{R}^{(p+1)\times(p+1)}$ appearing on the left-hand side of the normal equations (3) is a square, symmetric matrix known as the **Gram matrix**. Its (i, j)-th entry is the inner product of the i-th and j-th columns of X:

$$G_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i^\top \mathbf{x}_j.$$

The diagonal entries $G_{ii} = \|\mathbf{x}_i\|^2$ represent the squared norms of the feature vectors, while the off-diagonal entries encode the pairwise inner products between distinct features. In statistical terms, when the columns of X are centered, these off-diagonal entries are proportional to the sample covariances between features.

We first establish two fundamental lemmas that connect the concepts of linear independence and matrix invertibility to the structure of null spaces.

Lemma 4.1. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix with columns $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^m$. The columns of A are linearly independent if and only if* $\mathrm{Null}(A) = \{\mathbf{0}\}$.

Proof. Suppose first that the columns of A are linearly independent. Let $\mathbf{v} = (v_1, v_2, \ldots, v_n)^\top \in \mathrm{Null}(A)$, so that $A\mathbf{v} = \mathbf{0}$. By the definition of matrix-vector multiplication, this equation can be written as

$$A\mathbf{v} = v_1\mathbf{a}_1 + v_2\mathbf{a}_2 + \cdots + v_n\mathbf{a}_n = \mathbf{0}.$$

This is a linear combination of the columns of A that equals the zero vector. Since the columns are linearly independent by assumption, the only such linear combination is the trivial one, which requires $v_1 = v_2 = \cdots = v_n = 0$. Hence $\mathbf{v} = \mathbf{0}$, and since $\mathbf{v}$ was an arbitrary element of $\mathrm{Null}(A)$, we conclude that $\mathrm{Null}(A) = \{\mathbf{0}\}$.

Conversely, suppose that $\mathrm{Null}(A) = \{\mathbf{0}\}$. Consider an arbitrary linear combination of the columns of A that equals the zero vector:

$$c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n = \mathbf{0}.$$

Defining $\mathbf{c} = (c_1, c_2, \ldots, c_n)^\top$, this equation is equivalent to $A\mathbf{c} = \mathbf{0}$, which means $\mathbf{c} \in \mathrm{Null}(A)$. Since $\mathrm{Null}(A) = \{\mathbf{0}\}$ by assumption, we must have $\mathbf{c} = \mathbf{0}$, and therefore $c_1 = c_2 = \cdots = c_n = 0$. This demonstrates that the only linear combination of the columns that yields the zero vector is the trivial one, which is precisely the definition of linear independence. □

Lemma 4.2. *Let $A \in \mathbb{R}^{n\times n}$ be a square matrix. Then A is invertible if and only if* $\mathrm{Null}(A) = \{\mathbf{0}\}$.

Proof. Suppose first that A is invertible, and let $\mathbf{v} \in \mathrm{Null}(A)$, so that $A\mathbf{v} = \mathbf{0}$. Since A is invertible, we may left-multiply both sides by A^{-1} to obtain $A^{-1}(A\mathbf{v}) = A^{-1}\mathbf{0}$. The left-hand side simplifies to $(A^{-1}A)\mathbf{v} = I_n\mathbf{v} = \mathbf{v}$, while the right-hand side equals $\mathbf{0}$. Hence $\mathbf{v} = \mathbf{0}$, and since $\mathbf{v}$ was arbitrary, $\mathrm{Null}(A) = \{\mathbf{0}\}$.

Conversely, suppose that $\text{Null}(A) = \{\mathbf{0}\}$. By Theorem 4.1, this implies that the columns of A are linearly independent. Since A is an $n \times n$ matrix, it has exactly n columns. A set of n linearly independent vectors in $\mathbb{R}^n$ forms a basis for $\mathbb{R}^n$, and therefore the columns of A span all of $\mathbb{R}^n$. This means that for any $\mathbf{b} \in \mathbb{R}^n$, the equation $A\mathbf{x} = \mathbf{b}$ has at least one solution. Furthermore, the solution is unique: if $A\mathbf{x}_1 = \mathbf{b}$ and $A\mathbf{x}_2 = \mathbf{b}$, then $A(\mathbf{x}_1 - \mathbf{x}_2) = \mathbf{0}$, which implies $\mathbf{x}_1 - \mathbf{x}_2 \in \text{Null}(A) = \{\mathbf{0}\}$, and hence $\mathbf{x}_1 = \mathbf{x}_2$.

We now construct the inverse explicitly. For each standard basis vector $\mathbf{e}_j \in \mathbb{R}^n$ (for $j = 1, 2, \ldots, n$), let $\mathbf{b}_j$ denote the unique solution to $A\mathbf{b}_j = \mathbf{e}_j$. Define the matrix $B \in \mathbb{R}^{n \times n}$ by setting its j-th column equal to $\mathbf{b}_j$. Then

$$AB = [\, A\mathbf{b}_1 \mid A\mathbf{b}_2 \mid \cdots \mid A\mathbf{b}_n \,] = [\, \mathbf{e}_1 \mid \mathbf{e}_2 \mid \cdots \mid \mathbf{e}_n \,] = I_n.$$

A standard result in linear algebra states that if $AB = I_n$ for square matrices, then also $BA = I_n$. Hence $B = A^{-1}$, confirming that A is invertible. □

With these two lemmas established, we now solve the normal equation (3) for $\hat{\boldsymbol{\beta}}$, we must be able to invert the Gram matrix. The following theorem establishes the precise condition under which this is possible.

Theorem 4.3. *The Gram matrix $X^\top X$ is invertible if and only if the columns of the matrix X are linearly independent.*

Proof. By Lemma 4.2 and Lemma 4.1, the theorem therefore reduces to establishing the equality

$$\text{Null}(X^\top X) = \text{Null}(X).$$

We prove this equality by showing inclusion in both directions. To show that $\text{Null}(X) \subseteq \text{Null}(X^\top X)$, let $\mathbf{v}$ be an arbitrary vector satisfying $X\mathbf{v} = \mathbf{0}$. Left-multiplying both sides by $X^\top$ yields $(X^\top X)\mathbf{v} = X^\top \mathbf{0} = \mathbf{0}$, and hence $\mathbf{v} \in \text{Null}(X^\top X)$.

To show the reverse inclusion $\text{Null}(X^\top X) \subseteq \text{Null}(X)$, let $\mathbf{v}$ be an arbitrary vector satisfying $(X^\top X)\mathbf{v} = \mathbf{0}$. Left-multiplying both sides by $\mathbf{v}^\top$ gives $\mathbf{v}^\top (X^\top X)\mathbf{v} = 0$. Recognizing that $\mathbf{v}^\top X^\top = (X\mathbf{v})^\top$, we rewrite this as

$$(X\mathbf{v})^\top (X\mathbf{v}) = \|X\mathbf{v}\|_2^2 = 0.$$

By positive definiteness of the Euclidean norm, $\|X\mathbf{v}\|_2^2 = 0$ implies $X\mathbf{v} = \mathbf{0}$, and hence $\mathbf{v} \in \text{Null}(X)$.

Having established both inclusions, we conclude that $\text{Null}(X^\top X) = \text{Null}(X)$. □

The condition of linear independence of the columns of X is the mathematical embodiment of the statistical assumption of *no perfect multicollinearity*. When this condition holds, no feature can be expressed as an exact linear combination of the other features, and the regression problem is well-posed. Geometrically, linear independence ensures that the column vectors form a basis for $\text{Col}(X)$, and consequently any vector in that subspace—including the projection $\hat{\mathbf{y}}$—has a unique representation as a linear combination of the columns. If the columns were linearly dependent, infinitely many coefficient vectors $\hat{\boldsymbol{\beta}}$ would produce the same projection $\hat{\mathbf{y}}$, and the problem would lack a unique solution.

4.3 The Ordinary Least Squares Estimator

Assuming the columns of X are linearly independent, the Gram matrix $X^\top X$ is invertible. We can therefore solve the Normal Equations for $\hat{\boldsymbol{\beta}}$ by left-multiplying both sides by $(X^\top X)^{-1}$, i.e.,

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

Since $(X^\top X)^{-1}(X^\top X) = I_{p+1}$, this simplifies to

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}. \tag{4}$$

This closed-form expression is the **Ordinary Least Squares (OLS) estimator**, and it constitutes the cornerstone of linear regression analysis. The formula provides an explicit, computable solution for the coefficient vector that produces the orthogonal projection of $\mathbf{y}$ onto the column space of X.

Example 4.4. Consider a simple linear regression problem with $n = 3$ observations and a single predictor variable. Suppose we observe the data points $(x_1, y_1) = (1, 2)$, $(x_2, y_2) = (2, 3)$, and $(x_3, y_3) = (3, 5)$. The design matrix, augmented with an intercept column, and the response vector are

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix}.$$

We compute the OLS estimator $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}$ step by step. First, the Gram matrix and its inverse:

$$X^\top X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}, \quad (X^\top X)^{-1} = \frac{1}{6} \begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}.$$

Next, we compute $X^\top \mathbf{y}$:

$$X^\top \mathbf{y} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix} = \begin{pmatrix} 10 \\ 23 \end{pmatrix}.$$

The OLS estimator is therefore

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y} = \frac{1}{6} \begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix} \begin{pmatrix} 10 \\ 23 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 140 - 138 \\ -60 + 69 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 2 \\ 9 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 3/2 \end{pmatrix}.$$

Thus $\hat{\beta}_0 = \dfrac{1}{3}$ and $\hat{\beta}_1 = \dfrac{3}{2}$, yielding the fitted model

$$\hat{y} = \frac{1}{3} + \frac{3}{2} x.$$

The fitted values are

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 1/3 \\ 3/2 \end{pmatrix} = \begin{pmatrix} 11/6 \\ 10/3 \\ 29/6 \end{pmatrix} \approx \begin{pmatrix} 1.83 \\ 3.33 \\ 4.83 \end{pmatrix}.$$

The residual vector is $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (1/6, -1/3, 1/6)^\top$. One can verify that $\sum_{i=1}^{3} e_i = 0$ and $\sum_{i=1}^{3} x_i e_i = 0$, confirming the orthogonality conditions $X^\top \mathbf{e} = \mathbf{0}$.

Example 4.5. Consider a multiple regression problem with $n = 4$ observations and $p = 2$ predictor variables. Suppose we wish to model a response variable using two features, with the observed data given by

i	x_{i1}	x_{i2}	y_i
1	1	2	3
2	2	1	4
3	3	3	7
4	2	2	5

The design matrix, augmented with a leading column of ones for the intercept, and the response vector are

$$X = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 3 & 3 \\ 1 & 2 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 3 \\ 4 \\ 7 \\ 5 \end{pmatrix}.$$

We compute the OLS estimator $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}$. The Gram matrix is

$$X^\top X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 2 \\ 2 & 1 & 3 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 3 & 3 \\ 1 & 2 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 8 & 8 \\ 8 & 18 & 16 \\ 8 & 16 & 18 \end{pmatrix}.$$

Computing the inverse

$$(X^\top X)^{-1} = \frac{1}{8} \begin{pmatrix} 26 & -8 & -8 \\ -8 & 4 & 0 \\ -8 & 0 & 4 \end{pmatrix}.$$

Next, we compute

$$X^\top \mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 2 \\ 2 & 1 & 3 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 4 \\ 7 \\ 5 \end{pmatrix} = \begin{pmatrix} 19 \\ 40 \\ 39 \end{pmatrix}.$$

The OLS estimator is therefore

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y} = \frac{1}{8} \begin{pmatrix} 26 & -8 & -8 \\ -8 & 4 & 0 \\ -8 & 0 & 4 \end{pmatrix} \begin{pmatrix} 19 \\ 40 \\ 39 \end{pmatrix} = \frac{1}{8} \begin{pmatrix} 494 - 320 - 312 \\ -152 + 160 + 0 \\ -152 + 0 + 156 \end{pmatrix} = \frac{1}{8} \begin{pmatrix} -138 \\ 8 \\ 4 \end{pmatrix}.$$

Thus $\hat{\beta}_0 = -\dfrac{138}{8}$, $\hat{\beta}_1 = 1$, and $\hat{\beta}_2 = \dfrac{1}{2}$, yielding the fitted model

$$\hat{y} = -\frac{138}{8} + x_1 + \frac{1}{2} x_2$$

The fitted values and residuals can be computed as $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$ and $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. The orthogonality conditions $X^\top \mathbf{e} = \mathbf{0}$ guarantee that the residuals sum to zero and are uncorrelated with each predictor.

4.4 The Projection Matrix

We now construct the matrix operator that performs the orthogonal projection onto $\mathrm{Col}(X)$. Starting from the definition $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$ and substituting the derived expression for $\hat{\boldsymbol{\beta}}$, we obtain

$$\hat{\mathbf{y}} = X(X^\top X)^{-1} X^\top \mathbf{y}.$$

By associativity of matrix multiplication, we can isolate the matrices operating on $\mathbf{y}$ as follows

$$\hat{\mathbf{y}} = \underbrace{X(X^\top X)^{-1} X^\top}_{P} \mathbf{y} = P\mathbf{y}.$$

The matrix $P = X(X^\top X)^{-1}X^\top \in \mathbb{R}^{n\times n}$ is the **projection matrix**, often called the **hat matrix** because it is the operator that *puts a hat on* $\mathbf{y}$—that is, it transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$. This matrix projects any vector in $\mathbb{R}^n$ orthogonally onto the subspace $\mathrm{Col}(X)$.

To formally verify that P is an orthogonal projection matrix, we must establish two algebraic properties: symmetry ($P^\top = P$) and idempotence ($P^2 = P$).

Theorem 4.6. *The hat matrix $P = X(X^\top X)^{-1}X^\top$ is symmetric.*

Proof. Taking the transpose of P and applying the reversal property of transposes yields

$$P^\top = \left(X(X^\top X)^{-1}X^\top\right)^\top = (X^\top)^\top \left((X^\top X)^{-1}\right)^\top X^\top = X\left((X^\top X)^\top\right)^{-1}X^\top,$$

where we have used the fact that $(A^{-1})^\top = (A^\top)^{-1}$ for any invertible matrix A. Since the Gram matrix is symmetric, $(X^\top X)^\top = X^\top X$, and therefore

$$P^\top = X(X^\top X)^{-1}X^\top = P.$$

□

Theorem 4.7. *The hat matrix $P = X(X^\top X)^{-1}X^\top$ is idempotent, i.e., $P^2 = P$.*

Proof. Computing the square of P directly, we have

$$P^2 = \left(X(X^\top X)^{-1}X^\top\right)\left(X(X^\top X)^{-1}X^\top\right) = X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top.$$

Since $(X^\top X)^{-1}(X^\top X) = I_{p+1}$, this simplifies to

$$P^2 = X \cdot I_{p+1} \cdot (X^\top X)^{-1}X^\top = X(X^\top X)^{-1}X^\top = P.$$

□

The idempotence property has a natural geometric interpretation: projecting a vector that already lies in $\mathrm{Col}(X)$ leaves it unchanged. Since $\hat{\mathbf{y}} = P\mathbf{y}$ is in the column space, applying P again yields $P\hat{\mathbf{y}} = P^2\mathbf{y} = P\mathbf{y} = \hat{\mathbf{y}}$. The properties of symmetry and idempotence together constitute the defining algebraic characteristics of an orthogonal projection operator, providing a complete verification of the geometric function of the hat matrix.

Theorem 4.8. *Let $X \in \mathbb{R}^{n\times(p+1)}$ be a design matrix with linearly independent columns, and let $P = X(X^\top X)^{-1}X^\top$ be the associated hat matrix. Then the trace of P equals the number of columns of X, i.e.,*

$$\mathrm{tr}(P) = p + 1.$$

Proof. The proof relies on the cyclic property of the trace, which states that for any matrices A, B, and C of compatible dimensions, $\mathrm{tr}(ABC) = \mathrm{tr}(CAB) = \mathrm{tr}(BCA)$. We first establish this property before applying it to the hat matrix.

Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times m}$. The trace of $AB \in \mathbb{R}^{m\times m}$ is the sum of its diagonal entries:

$$\mathrm{tr}(AB) = \sum_{i=1}^{m}(AB)_{ii} = \sum_{i=1}^{m}\sum_{k=1}^{n} A_{ik}B_{ki}.$$

Similarly, the trace of $BA \in \mathbb{R}^{n\times n}$ is

$$\operatorname{tr}(BA) = \sum_{k=1}^{n}(BA)_{kk} = \sum_{k=1}^{n}\sum_{i=1}^{m} B_{ki}A_{ik}.$$

Since both expressions involve the same double sum (merely with the order of summation interchanged), we conclude that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ whenever both products are defined. Applying this property to the product $A = X(X^\top X)^{-1}$ and $B = X^\top$, we obtain $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, which gives

$$\operatorname{tr}\big(X(X^\top X)^{-1}X^\top\big) = \operatorname{tr}\big(X^\top X(X^\top X)^{-1}\big).$$

The expression on the right-hand side simplifies directly. Since $(X^\top X)^{-1}$ is the inverse of $X^\top X$, we have

$$X^\top X(X^\top X)^{-1} = I_{p+1},$$

where I_{p+1} denotes the $(p+1)\times(p+1)$ identity matrix. The trace of the identity matrix is simply the sum of its diagonal entries, each of which equals one:

$$\operatorname{tr}(I_{p+1}) = \sum_{j=1}^{p+1} 1 = p+1.$$

Combining these results, we conclude that

$$\operatorname{tr}(P) = \operatorname{tr}\big(X(X^\top X)^{-1}X^\top\big) = \operatorname{tr}(I_{p+1}) = p+1.$$

□

4.5 The Residual Maker Matrix

Complementary to the hat matrix is the **residual maker matrix**, defined as

$$M = I_n - P.$$

This matrix extracts the residual vector from the response:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - P\mathbf{y} = (I_n - P)\mathbf{y} = M\mathbf{y}.$$

The matrix M projects any vector in $\mathbb{R}^n$ onto the orthogonal complement of $\operatorname{Col}(X)$, which by Theorem (3.16) is $\operatorname{Col}(X)^\perp = \operatorname{Null}(X^\top)$. It is straightforward to verify that M is also symmetric and idempotent, confirming its status as an orthogonal projection matrix. Furthermore, $PM = MP = O$ (the zero matrix), reflecting the orthogonality of the two subspaces onto which P and M project.

Example 4.9. Consider a simple linear regression with $n = 3$ observations. The design matrix with intercept and the response vector are

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix}.$$

From our earlier computation, we have

$$X^\top X = \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}, \quad (X^\top X)^{-1} = \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}.$$

The hat matrix $P = X(X^\top X)^{-1}X^\top$ is computed as follows. First,

$$X(X^\top X)^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \cdot \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 8 & -3 \\ 2 & 0 \\ -4 & 3 \end{pmatrix}.$$

Then, multiplying by $X^\top$:

$$P = \frac{1}{6}\begin{pmatrix} 8 & -3 \\ 2 & 0 \\ -4 & 3 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix}.$$

We verify the defining properties. Symmetry is evident from inspection: $P^\top = P$. For idempotence, one can compute P^2 directly:

$$P^2 = \frac{1}{36}\begin{pmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix}\begin{pmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix} = \frac{1}{36}\begin{pmatrix} 30 & 12 & -6 \\ 12 & 12 & 12 \\ -6 & 12 & 30 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix} = P.$$

Applying the hat matrix to $\mathbf{y}$ yields the fitted values:

$$\hat{\mathbf{y}} = P\mathbf{y} = \frac{1}{6}\begin{pmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix}\begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 11 \\ 20 \\ 29 \end{pmatrix} = \begin{pmatrix} 11/6 \\ 10/3 \\ 29/6 \end{pmatrix}.$$

The residual maker matrix is $M = I_3 - P$:

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{6}\begin{pmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{pmatrix}.$$

Observe that every row of M is proportional to $(1, -2, 1)$, confirming that $\mathrm{Col}(M)$ is one-dimensional. The residuals are

$$\mathbf{e} = M\mathbf{y} = \frac{1}{6}\begin{pmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/6 \\ -1/3 \\ 1/6 \end{pmatrix}.$$

Note that $\mathbf{e}$ is indeed a scalar multiple of $(1, -2, 1)^\top$, which spans $\mathrm{Col}(X)^\perp$.

Example 4.10. Consider a regression problem with $n = 4$ observations and a single predictor, where the design matrix and response vector are

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 5 \end{pmatrix}.$$

Computing the components of the hat matrix:

$$X^\top X = \begin{pmatrix} 4 & 4 \\ 4 & 6 \end{pmatrix}, \quad (X^\top X)^{-1} = \frac{1}{8}\begin{pmatrix} 6 & -4 \\ -4 & 4 \end{pmatrix}.$$

The hat matrix is

$$P = X(X^\top X)^{-1}X^\top = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \cdot \frac{1}{8}\begin{pmatrix} 6 & -4 \\ -4 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}.$$

Carrying out the multiplication:

$$X(X^\top X)^{-1} = \frac{1}{8}\begin{pmatrix} 6 & -4 \\ 2 & 0 \\ 2 & 0 \\ -2 & 4 \end{pmatrix}, \quad P = \frac{1}{8}\begin{pmatrix} 6 & 2 & 2 & -2 \\ 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ -2 & 2 & 2 & 6 \end{pmatrix}.$$

The diagonal entries of P, denoted $h_{ii} = P_{ii}$, are called the *leverages*. They measure how much influence each observation has on its own fitted value. Here, $h_{11} = h_{44} = 6/8 = 0.75$ and $h_{22} = h_{33} = 2/8 = 0.25$. The observations at the extreme x-values ($x = 0$ and $x = 2$) have higher leverage than the central observations (both at $x = 1$), reflecting their greater influence on the fitted line.

The trace of the hat matrix equals the number of parameters

$$\operatorname{tr}(P) = \frac{6+2+2+6}{8} = \frac{16}{8} = 2 = p + 1.$$

The fitted values and residuals are

$$\hat{\mathbf{y}} = P\mathbf{y} = \frac{1}{8}\begin{pmatrix} 6 & 2 & 2 & -2 \\ 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ -2 & 2 & 2 & 6 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \\ 3 \\ 5 \end{pmatrix} = \frac{1}{8}\begin{pmatrix} 10 \\ 22 \\ 22 \\ 34 \end{pmatrix} = \begin{pmatrix} 5/4 \\ 11/4 \\ 11/4 \\ 17/4 \end{pmatrix},$$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} -1/4 \\ -3/4 \\ 1/4 \\ 3/4 \end{pmatrix}.$$

One can verify that $P\mathbf{e} = \mathbf{0}$ (the projection of the residual onto the column space vanishes) and $M\hat{\mathbf{y}} = \mathbf{0}$ (the fitted values have no component in the orthogonal complement), confirming the orthogonal decomposition $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$ with $\hat{\mathbf{y}} \perp \mathbf{e}$.

5 Minimization via Calculus

We now pivot from the geometric framework to the traditional analytical approach to demonstrate their profound equivalence. The analytical goal of linear regression is to find the coefficient vector β that minimizes the **Sum of Squared Residuals (SSR)**, also known as the Residual Sum of Squares (RSS).

5.1 The Least Squares Objective Function

The SSR is defined as the sum of the squared differences between the observed values y_i and the predicted values $\hat{y}_i = \mathbf{x}_i^T\beta$. This gives the scalar-valued loss function: We can immediately connect

this analytical objective to our geometric framework. This sum is precisely the squared Euclidean norm of the residual vector $\mathbf{e} = \mathbf{y} - X\beta$, i.e.,

$$L(\beta) = \|\mathbf{y} - X\beta\|_2^2. \tag{5}$$

Thus, the analytical goal of minimizing the sum of squared errors is identical to the geometric goal of finding the shortest possible residual vector—the very problem we solved using projection.

5.2 Minimization via the Gradient

To find the minimum of the multivariable function $L(\beta)$, we must compute its gradient with respect to the vector β and set the resulting vector of partial derivatives to the zero vector, i.e.,

$$\nabla_\beta L(\beta) = \mathbf{0}.$$

First, we expand the loss function using its vector transpose form:

$$L(\beta) = (\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\beta - \beta^T X^T\mathbf{y} + \beta^T X^T X\beta.$$

Since $\beta^T X^T\mathbf{y}$ is a scalar, it is equal to its own transpose, i.e.,

$$(\beta^T X^T\mathbf{y})^T = \mathbf{y}^T(X^T)^T(\beta^T)^T = \mathbf{y}^T X\beta.$$

The two middle terms are therefore identical. The loss function becomes

$$L(\beta) = \mathbf{y}^T\mathbf{y} - 2\beta^T X^T\mathbf{y} + \beta^T X^T X\beta.$$

We now compute the gradient using two standard matrix calculus identities $\nabla_{\mathbf{x}}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}$ and $\nabla_{\mathbf{x}}(\mathbf{x}^T A\mathbf{x}) = 2A\mathbf{x}$ for a symmetric matrix A. The gradient of $L(\beta)$ with respect to β is

$$\begin{aligned}\nabla_\beta L(\beta) &= \nabla_\beta(\mathbf{y}^T\mathbf{y} - 2\beta^T X^T\mathbf{y} + \beta^T X^T X\beta) \\ &= \mathbf{0} - 2X^T\mathbf{y} + 2(X^T X)\beta.\end{aligned}$$

Setting this gradient to the zero vector to find the critical point gives

$$-2X^T\mathbf{y} + 2X^T X\beta = \mathbf{0}.$$

Dividing by 2 and rearranging, we arrive at

$$X^T X\beta = X^T\mathbf{y}.$$

This is a remarkable result. The normal equation (3), which we derived from the purely geometric condition of orthogonality, have emerged again as the solution to the analytical problem of minimizing a squared-error loss function. This is not a coincidence; it reveals a deep and elegant unity between geometry and analysis. The squared Euclidean distance is the fundamental metric in both frameworks. The geometric approach finds the point of minimum distance by exploiting the properties of orthogonality. The calculus approach finds the point of minimum distance by finding where the function's rate of change is zero. Their convergence on the identical solution demonstrates that the geometric concept of *perpendicular* is the analytical equivalent of *local extremum* for quadratic forms defined by squared distances. This tells us that *least squares* is not an arbitrary choice of loss function; it is the unique loss function that corresponds to our intuitive geometric understanding of projection in Euclidean space.

To confirm that this solution corresponds to a minimum, one would compute the Hessian matrix (the matrix of second partial derivatives),

$$H = \nabla^2_\beta L(\beta) = 2X^T X.$$

As we have shown, the Gram matrix $X^T X$ is positive definite when the columns of X are linearly independent. A positive definite Hessian confirms that the critical point we found is indeed a unique global minimum.

5.3 The Convex Optimization Perspective

The analysis above confirmed that the critical point is a global minimum by examining the Hessian matrix. We now provide an alternative and more general perspective through the lens of convex optimization, which offers deeper insight into the structure of the least squares problem.

Definition 5.1 (Convex Function)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **convex** if for all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$ and all $\lambda \in [0, 1]$,

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{z}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{z}). \tag{6}$$

A function is **strictly convex** if the inequality is strict whenever $\mathbf{x} \neq \mathbf{z}$ and $\lambda \in (0, 1)$.

Geometrically, the condition (6) states that the line segment connecting any two points on the graph of f lies on or above the graph itself.

Example 5.2. Consider the univariate function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$. We verify that f is strictly convex by directly checking the defining inequality. Let $x, z \in \mathbb{R}$ with $x \neq z$, and let $\lambda \in (0, 1)$. The left-hand side of (6) is

$$f(\lambda x + (1 - \lambda)z) = (\lambda x + (1 - \lambda)z)^2 = \lambda^2 x^2 + 2\lambda(1 - \lambda)xz + (1 - \lambda)^2 z^2.$$

The right-hand side is

$$\lambda f(x) + (1 - \lambda) f(z) = \lambda x^2 + (1 - \lambda)z^2.$$

Computing the difference (right-hand side minus left-hand side):

$$\begin{aligned}
&\lambda x^2 + (1 - \lambda)z^2 - \lambda^2 x^2 - 2\lambda(1 - \lambda)xz - (1 - \lambda)^2 z^2 \\
&= \lambda(1 - \lambda)x^2 - 2\lambda(1 - \lambda)xz + \lambda(1 - \lambda)z^2 \\
&= \lambda(1 - \lambda)(x^2 - 2xz + z^2) \\
&= \lambda(1 - \lambda)(x - z)^2.
\end{aligned}$$

Since $\lambda \in (0, 1)$ implies $\lambda(1 - \lambda) > 0$, and since $x \neq z$ implies $(x - z)^2 > 0$, the difference is strictly positive. Therefore, $f(\lambda x + (1 - \lambda)z) < \lambda f(x) + (1 - \lambda) f(z)$, confirming that $f(x) = x^2$ is strictly convex.

Example 5.3. Consider the multivariate quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x},$$

where $A \in \mathbb{R}^{n \times n}$ is a symmetric positive semi-definite matrix. We show that f is convex. Let $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$ and $\lambda \in [0, 1]$. Define $\mathbf{w} = \lambda \mathbf{x} + (1 - \lambda)\mathbf{z}$. Expanding the left-hand side:

$$\begin{aligned}
f(\mathbf{w}) = \mathbf{w}^\top A \mathbf{w} &= (\lambda \mathbf{x} + (1 - \lambda)\mathbf{z})^\top A (\lambda \mathbf{x} + (1 - \lambda)\mathbf{z}) \\
&= \lambda^2 \mathbf{x}^\top A \mathbf{x} + 2\lambda(1 - \lambda)\mathbf{x}^\top A \mathbf{z} + (1 - \lambda)^2 \mathbf{z}^\top A \mathbf{z}.
\end{aligned}$$

The right-hand side of the convexity condition is

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{z}) = \lambda \mathbf{x}^\top A \mathbf{x} + (1-\lambda)\mathbf{z}^\top A \mathbf{z}.$$

Computing the difference as before:

$$\begin{aligned}
&\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{z}) - f(\mathbf{w}) \\
&= \lambda(1-\lambda)\mathbf{x}^\top A \mathbf{x} - 2\lambda(1-\lambda)\mathbf{x}^\top A \mathbf{z} + \lambda(1-\lambda)\mathbf{z}^\top A \mathbf{z} \\
&= \lambda(1-\lambda)(\mathbf{x}-\mathbf{z})^\top A (\mathbf{x}-\mathbf{z}).
\end{aligned}$$

Since A is positive semi-definite, $(\mathbf{x}-\mathbf{z})^\top A(\mathbf{x}-\mathbf{z}) \geq 0$ for all $\mathbf{x}, \mathbf{z}$. Combined with $\lambda(1-\lambda) \geq 0$ for $\lambda \in [0,1]$, the difference is non-negative, establishing convexity. If A is positive definite, the inequality is strict whenever $\mathbf{x} \neq \mathbf{z}$ and $\lambda \in (0,1)$, making f strictly convex.

The significance of convexity in optimization is captured by the following fundamental theorem.

Theorem 5.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then every local minimum of f is also a global minimum. Moreover, if f is strictly convex, then f has at most one global minimum.*

Proof. Suppose $\mathbf{x}^*$ is a local minimum of f but not a global minimum. Then there exists some $\mathbf{z} \in \mathbb{R}^n$ such that $f(\mathbf{z}) < f(\mathbf{x}^*)$. Consider the point $\mathbf{x}_\lambda = \lambda \mathbf{z} + (1-\lambda)\mathbf{x}^*$ for $\lambda \in (0,1)$. By convexity,

$$f(\mathbf{x}_\lambda) \leq \lambda f(\mathbf{z}) + (1-\lambda) f(\mathbf{x}^*) < \lambda f(\mathbf{x}^*) + (1-\lambda) f(\mathbf{x}^*) = f(\mathbf{x}^*).$$

As $\lambda \to 0$, the point $\mathbf{x}_\lambda$ approaches $\mathbf{x}^*$, yet $f(\mathbf{x}_\lambda) < f(\mathbf{x}^*)$. This contradicts the assumption that $\mathbf{x}^*$ is a local minimum. Hence every local minimum must be a global minimum.

For uniqueness under strict convexity, suppose $\mathbf{x}^*$ and $\mathbf{z}^*$ are both global minima with $\mathbf{x}^* \neq \mathbf{z}^*$, so that $f(\mathbf{x}^*) = f(\mathbf{z}^*) = m$ where m is the global minimum value. Consider $\mathbf{w} = \frac{1}{2}\mathbf{x}^* + \frac{1}{2}\mathbf{z}^*$. By strict convexity,

$$f(\mathbf{w}) < \frac{1}{2} f(\mathbf{x}^*) + \frac{1}{2} f(\mathbf{z}^*) = \frac{1}{2} m + \frac{1}{2} m = m.$$

This contradicts the fact that m is the global minimum. Therefore, the global minimum, if it exists, is unique. □

The connection between convexity and the Hessian matrix is provided by the following characterization.

Theorem 5.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Then f is convex if and only if its Hessian matrix $H(\mathbf{x}) = \nabla^2 f(\mathbf{x})$ is positive semi-definite for all $\mathbf{x} \in \mathbb{R}^n$. Furthermore, f is strictly convex if $H(\mathbf{x})$ is positive definite for all $\mathbf{x}$.*

We now apply these results to the least squares loss function $L(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$.

Theorem 5.6. *The least squares loss function* (5) *is convex. Moreover, L is strictly convex if and only if the columns of X are linearly independent.*

Proof. The Hessian of L is $H = \nabla^2_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = 2X^\top X$. This matrix is constant (independent of $\boldsymbol{\beta}$), so we need only verify its definiteness properties.

For any $\mathbf{v} \in \mathbb{R}^{p+1}$, we have

$$\mathbf{v}^\top (2X^\top X)\mathbf{v} = 2\mathbf{v}^\top X^\top X \mathbf{v} = 2(X\mathbf{v})^\top (X\mathbf{v}) = 2\|X\mathbf{v}\|_2^2 \geq 0.$$

Since $\|X\mathbf{v}\|_2^2 \geq 0$ for all $\mathbf{v}$, the Hessian is positive semi-definite, and hence L is convex by Theorem 5.5.

For strict convexity, the Hessian must be positive definite, meaning $\mathbf{v}^\top(X^\top X)\mathbf{v} > 0$ for all $\mathbf{v} \neq \mathbf{0}$. This is equivalent to requiring $\|X\mathbf{v}\|_2^2 > 0$ whenever $\mathbf{v} \neq \mathbf{0}$, which holds if and only if $X\mathbf{v} \neq \mathbf{0}$ for all nonzero $\mathbf{v}$. By Theorem 4.1, this condition is equivalent to the columns of X being linearly independent. □

This convex optimization perspective provides a powerful framework for understanding the least squares problem. When the columns of X are linearly independent, the loss function $L(\boldsymbol{\beta})$ is strictly convex, and by Theorem 5.4, the critical point $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1}X^\top \mathbf{y}$ is the unique global minimum. The strict convexity guarantees that no other coefficient vector can achieve an equally small or smaller loss.

When the columns of X are linearly dependent, the loss function remains convex but is no longer strictly convex. In this case, the set of global minima forms a convex set (in fact, an affine subspace), and infinitely many coefficient vectors achieve the same minimum loss. This corresponds to the geometric observation that multiple coefficient vectors can produce the same projection $\hat{\mathbf{y}}$ onto $\mathrm{Col}(X)$.

The convex optimization viewpoint thus unifies our geometric and analytical perspectives: the existence and uniqueness of the OLS estimator is fundamentally a consequence of the strict convexity of the squared Euclidean norm, which in turn reflects the positive definiteness of the inner product structure on $\mathbb{R}^n$.

6 Computational Realities and Numerical Stability

Thus far, our treatment has been purely theoretical. We have derived an elegant, closed-form solution for the OLS coefficients: $\hat{\beta} = (X^T X)^{-1}X^T \mathbf{y}$. For a mathematician, this might seem like the end of the story. For a data scientist or numerical analyst, however, this is where the story begins. The direct implementation of this formula in a computer program is fraught with peril due to the limitations of floating-point arithmetic.

6.1 Eigenvalues, Singular Values, and Matrix Norms

Before discussing the sensitivity of linear systems to perturbations, we introduce several fundamental concepts from matrix analysis that quantify the "size" of matrices and characterize their action on vectors.

Definition 6.1 (Eigenvalue and Eigenvector)**.** Let $A \in \mathbb{R}^{n\times n}$ be a square matrix. A scalar $\lambda \in \mathbb{C}$ is called an **eigenvalue** of A if there exists a nonzero vector $\mathbf{v} \in \mathbb{C}^n$ such that

$$A\mathbf{v} = \lambda\mathbf{v}.$$

The vector $\mathbf{v}$ is called an **eigenvector** of A corresponding to the eigenvalue λ. The set of all eigenvalues of A is called the **spectrum** of A, denoted $\sigma(A)$.

Example 6.2. Consider the symmetric matrix

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}.$$

To find the eigenvalues, we solve the characteristic equation $\det(A - \lambda I) = 0$:

$$\det \begin{pmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{pmatrix} = (3-\lambda)^2 - 1 = \lambda^2 - 6\lambda + 8 = (\lambda - 4)(\lambda - 2) = 0.$$

Thus the eigenvalues are $\lambda_1 = 4$ and $\lambda_2 = 2$. The corresponding eigenvectors are found by solving $(A - \lambda_i I)\mathbf{v} = \mathbf{0}$. For $\lambda_1 = 4$, we obtain $\mathbf{v}_1 = (1, 1)^\top$; for $\lambda_2 = 2$, we obtain $\mathbf{v}_2 = (1, -1)^\top$. One can verify that $A\mathbf{v}_1 = 4\mathbf{v}_1$ and $A\mathbf{v}_2 = 2\mathbf{v}_2$.

For symmetric matrices, eigenvalues are always real. However, to analyze general (possibly non-square) matrices, we require the more general notion of singular values.

Definition 6.3 (Singular Value Decomposition)**.** Let $A \in \mathbb{R}^{m \times n}$ be a matrix. The **singular value decomposition (SVD)** of A is a factorization of the form

$$A = U \Sigma V^\top,$$

where $U \in \mathbb{R}^{m \times m}$ is an orthogonal matrix (i.e., $U^\top U = I_m$), $V \in \mathbb{R}^{n \times n}$ is an orthogonal matrix, and $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with non-negative entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$ on its diagonal. The values σ_i are called the **singular values** of A, the columns of U are called the **left singular vectors**, and the columns of V are called the **right singular vectors**.

Theorem 6.4. *Every matrix $A \in \mathbb{R}^{m \times n}$ has a singular value decomposition. The singular values of A are uniquely determined and are equal to the non-negative square roots of the eigenvalues of $A^\top A$ (or equivalently, of $AA^\top$).*

Example 6.5. Consider the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1.0001 \end{pmatrix}.$$

To find the singular values, we compute the eigenvalues of $A^\top A$. Since A is symmetric, $A^\top A = A^2$:

$$A^\top A = A^2 = \begin{pmatrix} 2.0001 & 2.0001 \\ 2.0001 & 2.00020001 \end{pmatrix}.$$

The eigenvalues of $A^\top A$ are approximately $\lambda_1 \approx 4.0004$ and $\lambda_2 \approx 10^{-8}$. The singular values are the square roots: $\sigma_1 \approx 2.0001$ and $\sigma_2 \approx 0.0001$. For symmetric positive semi-definite matrices, the singular values coincide with the eigenvalues, which is the case here since A has positive eigenvalues.

We now introduce a way to measure the *size* of a matrix that is consistent with vector norms.

Definition 6.6 (Induced Matrix Norm)**.** Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^n$. The **induced matrix norm** (or **operator norm**) on $\mathbb{R}^{m \times n}$ is defined by

$$\|A\| = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\|\mathbf{x}\| = 1} \|A\mathbf{x}\|.$$

This quantity represents the maximum factor by which A can stretch any vector.

Definition 6.7 (Spectral Norm)**.** The **spectral norm** of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\|A\|_2$, is the induced matrix norm corresponding to the Euclidean vector norm. That is,

$$\|A\|_2 = \sup_{\|\mathbf{x}\|_2 = 1} \|A\mathbf{x}\|_2.$$

The following theorem provides a concrete formula for computing the spectral norm.

Theorem 6.8. *The spectral norm of a matrix $A \in \mathbb{R}^{m\times n}$ equals its largest singular value:*

$$\|A\|_2 = \sigma_{\max}(A).$$

Proof. Let $A = U\Sigma V^\top$ be the singular value decomposition of A, with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$, where $r = \min(m, n)$. For any unit vector $\mathbf{x}$ with $\|\mathbf{x}\|_2 = 1$, define $\mathbf{z} = V^\top \mathbf{x}$. Since V is orthogonal, $\|\mathbf{z}\|_2 = \|V^\top \mathbf{x}\|_2 = \|\mathbf{x}\|_2 = 1$. We compute

$$\|A\mathbf{x}\|_2^2 = \|U\Sigma V^\top \mathbf{x}\|_2^2 = \|\Sigma \mathbf{z}\|_2^2 = \sum_{i=1}^{r} \sigma_i^2 z_i^2 \leq \sigma_1^2 \sum_{i=1}^{r} z_i^2 \leq \sigma_1^2 \|\mathbf{z}\|_2^2 = \sigma_1^2.$$

Thus $\|A\mathbf{x}\|_2 \leq \sigma_1$ for all unit vectors $\mathbf{x}$. Equality is achieved when $\mathbf{x} = \mathbf{v}_1$, the first right singular vector, since then $\mathbf{z} = V^\top \mathbf{v}_1 = \mathbf{e}_1$ and $\|A\mathbf{v}_1\|_2 = \|\Sigma \mathbf{e}_1\|_2 = \sigma_1$. Therefore, $\|A\|_2 = \sigma_1 = \sigma_{\max}(A)$. □

Example 6.9. For the identity matrix I_n, all singular values equal 1, so $\|I_n\|_2 = 1$. For a diagonal matrix $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$, the singular values are $|d_1|, |d_2|, \ldots, |d_n|$, and hence $\|D\|_2 = \max_i |d_i|$.

These concepts allow us to express the condition number in terms of singular values.

Theorem 6.10. *For an invertible matrix $A \in \mathbb{R}^{n\times n}$, the condition number with respect to the spectral norm is*

$$\kappa(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}.$$

Proof. By Theorem 6.8, $\|A\|_2 = \sigma_{\max}(A)$. It remains to show that $\|A^{-1}\|_2 = 1/\sigma_{\min}(A)$.

Let $A = U\Sigma V^\top$ be the SVD of A. Since A is invertible, all singular values are positive, and we have $A^{-1} = V\Sigma^{-1}U^\top$, where $\Sigma^{-1} = \operatorname{diag}(1/\sigma_1, 1/\sigma_2, \ldots, 1/\sigma_n)$. The singular values of A^{-1} are therefore $1/\sigma_n, 1/\sigma_{n-1}, \ldots, 1/\sigma_1$ (in decreasing order, assuming $\sigma_1 \geq \cdots \geq \sigma_n > 0$). The largest singular value of A^{-1} is $1/\sigma_n = 1/\sigma_{\min}(A)$. Hence,

$$\|A^{-1}\|_2 = \sigma_{\max}(A^{-1}) = \frac{1}{\sigma_{\min}(A)}.$$

Combining these results:

$$\kappa(A) = \|A\|_2 \|A^{-1}\|_2 = \sigma_{\max}(A) \cdot \frac{1}{\sigma_{\min}(A)} = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}.$$

□

For symmetric positive definite matrices, there is a further simplification.

Corollary 6.11. *If $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$, then*

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} = \frac{\lambda_1}{\lambda_n}.$$

Proof. For a symmetric positive definite matrix, the singular values coincide with the eigenvalues (since all eigenvalues are positive). Therefore, $\sigma_{\max}(A) = \lambda_{\max}(A)$ and $\sigma_{\min}(A) = \lambda_{\min}(A)$, and the result follows from Theorem 6.10. □

This corollary is particularly relevant to linear regression, since the Gram matrix $X^\top X$ is symmetric and (when X has linearly independent columns) positive definite.

6.2 The Fragility of the Normal Equations

The theoretical perfection of the normal Equations belies their computational fragility. The issue lies not in the theory, but in the finite precision of computer arithmetic. To understand why, we must introduce the concept of the condition number of a matrix.

Definition 6.12 (Condition Number)**.** The **condition number** of an invertible matrix A, denoted $\kappa(A)$, is a measure of how sensitive the solution of the system $A\mathbf{x} = \mathbf{b}$ is to perturbations in A or $\mathbf{b}$. It is defined as:

$$\kappa(A) = \|A\|\|A^{-1}\|$$

where $\|\cdot\|$ is an induced matrix norm. A matrix with a large condition number is said to be **ill-conditioned**, while a matrix with a condition number close to 1 is **well-conditioned**.

Example 6.13. Consider the 2×2 identity matrix $A = I_2$. Since $A^{-1} = I_2$ as well, using the spectral norm (induced by the Euclidean vector norm), we have $\|I_2\| = 1$ and $\|I_2^{-1}\| = 1$. Therefore, the condition number is

$$\kappa(I_2) = \|I_2\| \cdot \|I_2^{-1}\| = 1 \cdot 1 = 1.$$

This is the smallest possible condition number, and the identity matrix is perfectly well-conditioned. Consider solving $I_2\mathbf{x} = \mathbf{b}$, which gives $\mathbf{x} = \mathbf{b}$. If we perturb $\mathbf{b}$ to $\mathbf{b} + \boldsymbol{\delta}$, the solution becomes $\mathbf{x} + \boldsymbol{\delta}$. The relative change in the solution exactly equals the relative change in the input: there is no amplification of errors. More generally, any orthogonal matrix Q (satisfying $Q^\top Q = I$) has condition number $\kappa(Q) = 1$, since $Q^{-1} = Q^\top$ and both Q and $Q^\top$ have spectral norm equal to 1.

Example 6.14. Consider the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1.0001 \end{pmatrix}.$$

This matrix is invertible, with

$$A^{-1} = \frac{1}{0.0001}\begin{pmatrix} 1.0001 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 10001 & -10000 \\ -10000 & 10000 \end{pmatrix}.$$

The eigenvalues of A are approximately $\lambda_1 \approx 2.0001$ and $\lambda_2 \approx 0.0001$. Using the spectral norm, which equals the largest singular value, the condition number is approximately

$$\kappa(A) = \frac{\sigma_{\max}}{\sigma_{\min}} \approx \frac{2.0001}{0.0001} \approx 20000.$$

This matrix is severely ill-conditioned. To illustrate the practical consequence, consider solving $A\mathbf{x} = \mathbf{b}$ with $\mathbf{b} = (2, 2.0001)^\top$. The exact solution is $\mathbf{x} = (1, 1)^\top$. Now suppose we perturb $\mathbf{b}$ slightly to $\tilde{\mathbf{b}} = (2, 2.0002)^\top$, a relative change of approximately 0.005%. Solving $A\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ yields $\tilde{\mathbf{x}} = (0, 2)^\top$. The relative change in the solution is $\|\tilde{\mathbf{x}} - \mathbf{x}\|/\|\mathbf{x}\| = \|(-1, 1)^\top\|/\|(1, 1)^\top\| = 1$, or 100%. A tiny perturbation in the input has been amplified by a factor on the order of $\kappa(A)$, dramatically altering the solution. In the context of linear regression, if the Gram matrix $X^\top X$ has a large condition number (due to near-collinearity among features), the computed OLS estimates $\hat{\boldsymbol{\beta}}$ may be highly unreliable.

An ill-conditioned matrix acts as an error amplifier. Small relative errors in the input data can lead to large relative errors in the output solution. The key problem with the Normal Equations approach is the explicit formation of the Gram matrix, $X^T X$. This single step can be a numerical

disaster, because of the following critical result: The condition number of the Gram matrix is the square of the condition number of the original design matrix.

The implications of this are severe. Suppose our design matrix X has some nearly collinear columns, a common occurrence in real-world datasets. This might make X moderately ill-conditioned, say $\kappa(X) \approx 10^7$. When we form X^TX, its condition number becomes $\kappa(X^TX) \approx (10^7)^2 = 10^{14}$. Standard double-precision floating-point arithmetic stores about 16 decimal digits of precision. A condition number of 10^{14} means that in the worst case, we could lose up to 14 of those 16 digits of precision during the computation of the inverse.[2] The resulting coefficient vector $\hat{\beta}$ could be completely meaningless, corrupted by catastrophic round-off error. The act of explicitly forming X^TX is therefore considered **numerically unstable** and is avoided in all serious numerical software.

6.3 QR Decomposition Apporach

Fortunately, there exist numerically superior methods for solving the least squares problem that completely bypass the formation of the Gram matrix $X^\top X$. The first and most widely used of these is based on the QR decomposition.

Definition 6.15 (QR Decomposition)**.** Let $X \in \mathbb{R}^{n\times m}$ be a matrix with $n \geq m$. A **QR decomposition** of X is a factorization of the form

$$X = QR,$$

where $Q \in \mathbb{R}^{n\times m}$ is a matrix with orthonormal columns (i.e., $Q^\top Q = I_m$), and $R \in \mathbb{R}^{m\times m}$ is an upper triangular matrix.

Theorem 6.16. *Every matrix $X \in \mathbb{R}^{n\times m}$ with $n \geq m$ admits a QR decomposition. If X has full column rank (i.e.,* $\operatorname{rank}(X) = m$*), then there exists a unique QR decomposition in which R has strictly positive diagonal entries.*

Geometrically, the QR decomposition can be viewed as a stable, matrix-based implementation of the Gram-Schmidt orthonormalization process. The columns of Q form an orthonormal basis for the column space of X, so that $\operatorname{Col}(Q) = \operatorname{Col}(X)$. The upper triangular matrix R encodes the coefficients expressing the original columns of X as linear combinations of this orthonormal basis.

Example 6.17. Consider the design matrix from our earlier simple linear regression example:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}.$$

Applying the Gram-Schmidt process to the columns of X yields the QR decomposition $X = QR$ with

$$Q = \frac{1}{\sqrt{6}} \begin{pmatrix} \sqrt{2} & -\sqrt{3} \\ \sqrt{2} & 0 \\ \sqrt{2} & \sqrt{3} \end{pmatrix} = \begin{pmatrix} 1/\sqrt{3} & -1/\sqrt{2} \\ 1/\sqrt{3} & 0 \\ 1/\sqrt{3} & 1/\sqrt{2} \end{pmatrix}, \quad R = \begin{pmatrix} \sqrt{3} & 2\sqrt{3} \\ 0 & \sqrt{2} \end{pmatrix}.$$

One can verify that $Q^\top Q = I_2$ and $QR = X$.

We now demonstrate how the QR decomposition provides a numerically stable approach to solving the least squares problem. Our goal is to find $\hat{\boldsymbol{\beta}}$ that minimizes $\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$. Substituting $X = QR$ into the objective function yields

$$\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 = \|\mathbf{y} - QR\boldsymbol{\beta}\|_2^2.$$

Since Q has orthonormal columns, it preserves the Euclidean inner product in the sense that $\|Q\mathbf{z}\|_2 = \|\mathbf{z}\|_2$ for any $\mathbf{z} \in \mathbb{R}^m$. Consequently, left-multiplying by $Q^\top$ is an isometry on the column space of Q, and we can write

$$\|\mathbf{y} - QR\boldsymbol{\beta}\|_2^2 = \|Q^\top\mathbf{y} - Q^\top QR\boldsymbol{\beta}\|_2^2 + \|(I_n - QQ^\top)\mathbf{y}\|_2^2.$$

The second term represents the component of $\mathbf{y}$ orthogonal to $\mathrm{Col}(X)$, which is independent of $\boldsymbol{\beta}$. Since $Q^\top Q = I_m$, the first term simplifies to $\|Q^\top\mathbf{y} - R\boldsymbol{\beta}\|_2^2$. Thus,

$$\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 = \|Q^\top\mathbf{y} - R\boldsymbol{\beta}\|_2^2 + \text{constant}.$$

The minimum is achieved when $R\boldsymbol{\beta} = Q^\top\mathbf{y}$, yielding the triangular system

$$R\hat{\boldsymbol{\beta}} = Q^\top\mathbf{y}.$$

This system is far better conditioned than the Normal Equations. The following theorem quantifies this advantage.

Theorem 6.18. *Let $X = QR$ be the QR decomposition of a full-rank matrix X. Then $\kappa(R) = \kappa(X)$.*

Proof. Since Q has orthonormal columns, $Q^\top Q = I_m$, and thus $\|Q\mathbf{z}\|_2 = \|\mathbf{z}\|_2$ for all $\mathbf{z} \in \mathbb{R}^m$. For any $\mathbf{z} \neq \mathbf{0}$,

$$\frac{\|X\mathbf{z}\|_2}{\|\mathbf{z}\|_2} = \frac{\|QR\mathbf{z}\|_2}{\|\mathbf{z}\|_2} = \frac{\|R\mathbf{z}\|_2}{\|\mathbf{z}\|_2}.$$

Taking the supremum over all nonzero $\mathbf{z}$ yields $\|X\|_2 = \|R\|_2$. Similarly, since $X^{-1} = R^{-1}Q^\top$ (in the sense of left inverses) and $Q^\top$ also preserves norms, we have $\|X^\dagger\|_2 = \|R^{-1}\|_2$, where $X^\dagger$ denotes the Moore-Penrose pseudoinverse. Therefore,

$$\kappa(R) = \|R\|_2\|R^{-1}\|_2 = \|X\|_2\|X^\dagger\|_2 = \kappa(X).$$

□

This theorem demonstrates that the QR approach avoids the catastrophic squaring of the condition number inherent in forming $X^\top X$. The condition number of the system we solve is $\kappa(R) = \kappa(X)$, rather than $\kappa(X^\top X) = \kappa(X)^2$.

Furthermore, because R is upper triangular, the system $R\hat{\boldsymbol{\beta}} = Q^\top\mathbf{y}$ can be solved very efficiently and accurately using a process called **back substitution**, without computing a matrix inverse. If we denote $\mathbf{c} = Q^\top\mathbf{y}$ and write the system explicitly as

$$\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ 0 & r_{22} & \cdots & r_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & r_{mm} \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_m \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{pmatrix},$$

then the solution proceeds from the bottom row upward:

$$\hat{\beta}_m = \frac{c_m}{r_{mm}}, \quad \hat{\beta}_{m-1} = \frac{c_{m-1} - r_{m-1,m}\hat{\beta}_m}{r_{m-1,m-1}}, \quad \ldots, \quad \hat{\beta}_j = \frac{c_j - \sum_{k=j+1}^{m} r_{jk}\hat{\beta}_k}{r_{jj}}.$$

Each step involves only scalar divisions and subtractions, requiring $O(m^2)$ operations in total, compared to $O(m^3)$ for a general matrix inversion.

Example 6.19. Continuing with the previous example, suppose $\mathbf{y} = (2, 3, 5)^\top$. We first compute

$$\mathbf{c} = Q^\top \mathbf{y} = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix} = \begin{pmatrix} 10/\sqrt{3} \\ 3/\sqrt{2} \end{pmatrix}.$$

The triangular system $R\hat{\boldsymbol{\beta}} = \mathbf{c}$ is

$$\begin{pmatrix} \sqrt{3} & 2\sqrt{3} \\ 0 & \sqrt{2} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 10/\sqrt{3} \\ 3/\sqrt{2} \end{pmatrix}.$$

Back substitution yields $\hat{\beta}_1 = (3/\sqrt{2})/\sqrt{2} = 3/2$ and $\hat{\beta}_0 = (10/\sqrt{3} - 2\sqrt{3} \cdot 3/2)/\sqrt{3} = (10/\sqrt{3} - 3\sqrt{3})/\sqrt{3} = 10/3 - 3 = 1/3$. Thus $\hat{\boldsymbol{\beta}} = (1/3, 3/2)^\top$, which agrees with our earlier computation using the Normal Equations.

6.4 Singular Value Decomposition Approach

While the QR decomposition represents a substantial improvement over the Normal Equations, the most powerful, insightful, and numerically robust tool for solving the least squares problem is the singular value decomposition (SVD), which we introduced in an earlier section. Recall from Theorem 6.4 that any matrix $X \in \mathbb{R}^{n \times m}$ admits a factorization $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{m \times m}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{n \times m}$ is a diagonal matrix containing the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, where $r = \operatorname{rank}(X)$.

The SVD provides a complete geometric decomposition of the linear transformation represented by X. Any such transformation can be decomposed into three fundamental operations: a rotation (or reflection) in the domain space ($V^\top$), a scaling along the principal axes (Σ), and a rotation (or reflection) in the codomain space (U). This decomposition reveals the intrinsic geometric structure of the matrix, independent of the choice of coordinates.

To apply the SVD to the least squares problem, we introduce the generalization of matrix inversion that applies to arbitrary matrices, including those that are rectangular or rank-deficient.

Definition 6.20 (Moore-Penrose Pseudoinverse)**.** Let $X \in \mathbb{R}^{n \times m}$ be a matrix with singular value decomposition $X = U\Sigma V^\top$. The **Moore-Penrose pseudoinverse** of X, denoted X^+, is defined as

$$X^+ = V\Sigma^+ U^\top,$$

where $\Sigma^+ \in \mathbb{R}^{m \times n}$ is obtained by transposing Σ and replacing each nonzero diagonal entry σ_i by its reciprocal $1/\sigma_i$.

Example 6.21. Consider a diagonal matrix $\Sigma \in \mathbb{R}^{3 \times 2}$ with singular values $\sigma_1 = 4$ and $\sigma_2 = 2$:

$$\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 2 \\ 0 & 0 \end{pmatrix}, \quad \Sigma^+ = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/2 & 0 \end{pmatrix}.$$

If one of the singular values were zero, say $\sigma_2 = 0$, then the corresponding entry in Σ^+ would also be zero, not undefined:

$$\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma^+ = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

This convention allows the pseudoinverse to be defined for all matrices, regardless of rank.

The pseudoinverse satisfies four characteristic properties, known as the Moore-Penrose conditions, which uniquely determine it.

Theorem 6.22 (Moore-Penrose Conditions)**.** *For any matrix $X \in \mathbb{R}^{n\times m}$, the pseudoinverse X^+ is the unique matrix satisfying:*

1. $XX^+X = X$
2. $X^+XX^+ = X^+$
3. $(XX^+)^\top = XX^+$
4. $(X^+X)^\top = X^+X$

The following theorem establishes the connection between the pseudoinverse and the least squares problem.

Theorem 6.23. *Let $X \in \mathbb{R}^{n\times m}$ and $\mathbf{y} \in \mathbb{R}^n$. The vector $\hat{\boldsymbol{\beta}} = X^+\mathbf{y}$ is a least squares solution, i.e., it minimizes $\|\mathbf{y} - X\boldsymbol{\beta}\|_2$. Moreover, among all least squares solutions, $X^+\mathbf{y}$ has the smallest Euclidean norm $\|\boldsymbol{\beta}\|_2$.*

Proof. Let $X = U\Sigma V^\top$ be the SVD of X, with $r = \operatorname{rank}(X)$ nonzero singular values. Partition $U = (\,U_1 \mid U_2\,)$ where $U_1 \in \mathbb{R}^{n\times r}$ contains the first r columns, and similarly $V = (\,V_1 \mid V_2\,)$ with $V_1 \in \mathbb{R}^{m\times r}$. Let $\Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$ be the nonzero part of Σ. Then $X = U_1\Sigma_r V_1^\top$.

For any $\boldsymbol{\beta} \in \mathbb{R}^m$, write $\boldsymbol{\beta} = V_1\boldsymbol{\alpha} + V_2\boldsymbol{\gamma}$ where $\boldsymbol{\alpha} \in \mathbb{R}^r$ and $\boldsymbol{\gamma} \in \mathbb{R}^{m-r}$. Then $X\boldsymbol{\beta} = U_1\Sigma_r\boldsymbol{\alpha}$, which is independent of $\boldsymbol{\gamma}$. The residual is

$$\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 = \|U_1^\top\mathbf{y} - \Sigma_r\boldsymbol{\alpha}\|_2^2 + \|U_2^\top\mathbf{y}\|_2^2.$$

The second term is constant; the first is minimized when $\boldsymbol{\alpha} = \Sigma_r^{-1}U_1^\top\mathbf{y}$. Any $\boldsymbol{\gamma}$ gives a least squares solution, but $\|\boldsymbol{\beta}\|_2^2 = \|\boldsymbol{\alpha}\|_2^2 + \|\boldsymbol{\gamma}\|_2^2$ is minimized when $\boldsymbol{\gamma} = \mathbf{0}$. The minimum-norm solution is therefore

$$\hat{\boldsymbol{\beta}} = V_1\Sigma_r^{-1}U_1^\top\mathbf{y} = V\Sigma^+U^\top\mathbf{y} = X^+\mathbf{y}.$$

□

When X has full column rank, the pseudoinverse reduces to the familiar expression from the Normal Equations.

Theorem 6.24. *If $X \in \mathbb{R}^{n\times m}$ has full column rank (i.e., $\operatorname{rank}(X) = m$), then*

$$X^+ = (X^\top X)^{-1}X^\top.$$

Proof. When X has full column rank, $X^\top X$ is invertible. We verify that $(X^\top X)^{-1}X^\top$ satisfies the four Moore-Penrose conditions. Let $A = (X^\top X)^{-1}X^\top$.

For condition (1): $XAX = X(X^\top X)^{-1}X^\top X = XI_m = X$.

For condition (2): $AXA = (X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = (X^\top X)^{-1}X^\top = A$.

For condition (3): $XA = X(X^\top X)^{-1}X^\top$, which is symmetric as shown in our analysis of the hat matrix.

For condition (4): $AX = (X^\top X)^{-1}X^\top X = I_m$, which is trivially symmetric.

By uniqueness of the pseudoinverse (Theorem 6.22), $X^+ = (X^\top X)^{-1}X^\top$. □

The SVD approach to least squares thus computes the OLS estimator as

$$\hat{\boldsymbol{\beta}} = X^{+}\mathbf{y} = V\Sigma^{+}U^{\top}\mathbf{y}.$$

This method offers several compelling advantages over the Normal Equations and even the QR decomposition.

1. **Numerical stability**: The SVD is computed using algorithms that are backward stable, and the condition number of the problem is determined by $\sigma_{\max}/\sigma_{\min}$, which is $\kappa(X)$ rather than $\kappa(X)^2$.

2. **Handling rank deficiency**: When X is rank-deficient (i.e., has linearly dependent columns), the Gram matrix $X^{\top}X$ is singular and the Normal Equations have no unique solution. The SVD gracefully handles this situation by setting the reciprocals of zero singular values to zero in Σ^{+}, yielding the unique minimum-norm least squares solution.

3. **Diagnostic information**: The singular values directly reveal the numerical rank of X and the sensitivity of the solution to perturbations. Small singular values indicate near-collinearity among features.

Example 6.25. Consider a rank-deficient design matrix where the third column is the sum of the first two:

$$X = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}.$$

The matrix X has rank 2, and consequently $X^{\top}X$ is singular:

$$X^{\top}X = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 2 & 3 \\ 3 & 3 & 6 \end{pmatrix}, \quad \det(X^{\top}X) = 0.$$

The Normal Equations $(X^{\top}X)\boldsymbol{\beta} = X^{\top}\mathbf{y}$ have infinitely many solutions. However, the SVD yields singular values $\sigma_1 \approx 3.86$, $\sigma_2 \approx 1.00$, and $\sigma_3 = 0$. The pseudoinverse X^{+} is computed by inverting only the nonzero singular values:

$$\Sigma^{+} = \begin{pmatrix} 1/\sigma_1 & 0 & 0 \\ 0 & 1/\sigma_2 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The minimum-norm solution $\hat{\boldsymbol{\beta}} = X^{+}\mathbf{y}$ is uniquely determined and represents the coefficient vector with smallest Euclidean norm among all vectors that minimize the residual sum of squares.

In practice, the QR decomposition is often preferred for well-conditioned, full-rank problems due to its computational efficiency, while the SVD is the method of choice when rank deficiency is suspected or when diagnostic information about the singular values is desired.

6.5 Geometric Interpretation of the SVD

The computational advantages of the SVD are substantial, but its true power for data science lies in the geometric insight it provides. The SVD furnishes the ultimate unification of the two geometric perspectives of the data matrix that we introduced at the beginning of this lecture.

Recall that we can view the data matrix $X \in \mathbb{R}^{n\times p}$ in two complementary ways: as a cloud of n points in the p-dimensional feature space (the row space perspective), or as a collection of p vectors in the n-dimensional observation space (the column space perspective). The SVD reveals the intrinsic geometric structure of both views simultaneously.

Theorem 6.26. *Let $X = U\Sigma V^\top$ be the singular value decomposition of $X \in \mathbb{R}^{n\times p}$ with $r = \operatorname{rank}(X)$. Write $U = (\,U_1 \mid U_2\,)$ where $U_1 \in \mathbb{R}^{n\times r}$ contains the first r columns, and similarly $V = (\,V_1 \mid V_2\,)$ with $V_1 \in \mathbb{R}^{p\times r}$. Then:*

1. *The columns of V_1 form an orthonormal basis for $\operatorname{Row}(X)$, the row space of X.*
2. *The columns of V_2 form an orthonormal basis for $\operatorname{Null}(X)$, the null space of X.*
3. *The columns of U_1 form an orthonormal basis for $\operatorname{Col}(X)$, the column space of X.*
4. *The columns of U_2 form an orthonormal basis for $\operatorname{Null}(X^\top)$, the left null space of X.*

This theorem demonstrates that the SVD simultaneously provides orthonormal bases for all four fundamental subspaces associated with X. We now examine the geometric significance of each component.

The right singular vectors, comprising the columns of V, provide an orthonormal basis for the row space of X. In the context of data analysis, where rows represent observations and columns represent features, these vectors define the principal axes of the data cloud in feature space. Specifically, $\mathbf{v}_1$ points in the direction along which the data exhibits maximum variance, $\mathbf{v}_2$ points in the direction of maximum variance orthogonal to $\mathbf{v}_1$, and so forth. The right singular vectors are precisely the eigenvectors of the Gram matrix $X^\top X$, and this observation forms the mathematical foundation of principal component analysis (PCA).

The left singular vectors, comprising the columns of U, provide an orthonormal basis for the column space of X. These vectors are the fundamental building blocks of the observation space as determined by the features. Each left singular vector $\mathbf{u}_j$ represents a pattern across all n observations that captures a distinct mode of variation in the data.

The relationship between these two perspectives is encoded in the SVD equation. Rewriting $X = U\Sigma V^\top$ as $XV = U\Sigma$, or equivalently $X\mathbf{v}_j = \sigma_j \mathbf{u}_j$ for each j, we obtain an explicit mathematical link between the two geometric views. This equation states that the principal directions of variance in feature space (the columns of V) are mapped by the linear transformation X into an orthogonal basis for the observation space (the columns of U), with the mapping scaled by the singular values. The singular value σ_j quantifies the "importance" of the j-th principal direction: larger singular values correspond to directions that capture more of the data's variance.

Example 6.27. Consider a data matrix $X \in \mathbb{R}^{100\times 3}$ representing 100 observations of 3 features. Suppose the SVD yields singular values $\sigma_1 = 50$, $\sigma_2 = 10$, and $\sigma_3 = 0.01$. The ratio $\sigma_1/\sigma_3 = 5000$ indicates that the data is nearly planar in feature space: the variation along $\mathbf{v}_3$ is negligible compared to the variation along $\mathbf{v}_1$. The data cloud is essentially two-dimensional, confined to the plane spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$. This insight, revealed directly by the singular values, would be obscured by examining the raw data or even the Gram matrix.

Example 6.28. Consider the full-rank matrix

$$X = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{3\times 2}.$$

The rank is $r = 2$, so the null space of X is trivial. The SVD yields

$$X = U\Sigma V^\top = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Partitioning as in the theorem, with U_1 containing the first two columns of U and U_2 containing the third:

$$U_1 = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 \end{pmatrix}, \quad U_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad V_1 = V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We verify the four subspaces. Since $r = p = 2$, the matrix V_2 is empty, confirming that $\mathrm{Null}(X) = \{\mathbf{0}\}$. The columns of $V_1 = I_2$ trivially span all of $\mathbb{R}^2$, which equals $\mathrm{Row}(X)$ since X has full column rank. The columns of U_1 span the column space: any vector in $\mathrm{Col}(X)$ has zero third component, which matches $\mathrm{span}\{(1/\sqrt{2}, 1/\sqrt{2}, 0)^\top, (1/\sqrt{2}, -1/\sqrt{2}, 0)^\top\}$. Finally, $U_2 = (0, 0, 1)^\top$ spans the left null space, and indeed $X^\top (0, 0, 1)^\top = \mathbf{0}$.

Example 6.29. Consider the rank-deficient matrix

$$X = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix} \in \mathbb{R}^{2\times 3}.$$

Since the second row is twice the first, we have $r = \mathrm{rank}(X) = 1$. The SVD is

$$X = U\Sigma V^\top = \begin{pmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ 2/\sqrt{5} & -1/\sqrt{5} \end{pmatrix} \begin{pmatrix} \sqrt{70} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{14} & 2/\sqrt{14} & 3/\sqrt{14} \\ -5/\sqrt{30} & 2/\sqrt{30} & 1/\sqrt{30} \\ 1/\sqrt{105} & -4/\sqrt{105} & 5/\sqrt{105} \end{pmatrix}.$$

Partitioning according to $r = 1$:

$$U_1 = \begin{pmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{pmatrix}, \quad U_2 = \begin{pmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{pmatrix},$$

$$V_1 = \begin{pmatrix} 1/\sqrt{14} \\ 2/\sqrt{14} \\ 3/\sqrt{14} \end{pmatrix}, \quad V_2 = \begin{pmatrix} -5/\sqrt{30} & 1/\sqrt{105} \\ 2/\sqrt{30} & -4/\sqrt{105} \\ 1/\sqrt{30} & 5/\sqrt{105} \end{pmatrix}.$$

We verify each subspace. The row space $\mathrm{Row}(X) = \mathrm{span}\{(1, 2, 3)\}$ is one-dimensional, and $V_1 = (1, 2, 3)^\top/\sqrt{14}$ is indeed a unit vector in this direction. The null space $\mathrm{Null}(X)$ is two-dimensional; the columns of V_2 are orthonormal vectors satisfying $XV_2 = O$, which can be verified by direct computation. The column space $\mathrm{Col}(X) = \mathrm{span}\{(1, 2)^\top\}$ is spanned by $U_1 = (1, 2)^\top/\sqrt{5}$. Finally, the left null space $\mathrm{Null}(X^\top)$ is spanned by $U_2 = (2, -1)^\top/\sqrt{5}$, and one can check that $X^\top U_2 = \mathbf{0}$. This example illustrates how the SVD reveals the complete geometric structure even when the matrix is rank-deficient.

The SVD is therefore not merely a computational algorithm for solving least squares problems; it is the fundamental geometric decomposition of the data matrix, revealing the intrinsic structure and relationships between the feature space and the observation space. This dual interpretation makes the SVD an indispensable tool for understanding high-dimensional data.

Table 1: Comparison of least squares solution methods.

	Normal Equations	**QR Decomposition**	**SVD**
Key equation	$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}$	$R\hat{\boldsymbol{\beta}} = Q^\top \mathbf{y}$	$\hat{\boldsymbol{\beta}} = V\Sigma^+ U^\top \mathbf{y}$
Condition number	$\kappa(X)^2$	$\kappa(X)$	$\kappa(X)$
Rank deficiency	Not handled	Not handled	Handled
Computational cost	$O(pm^2 + m^3)$	$O(pm^2)$	$O(pm^2 + m^3)$
Geometric insight	Minimal	Moderate	Comprehensive

6.6 Summary of Least Squares Solution Methods

We conclude this section by comparing the three methods for solving the least squares problem. Table 1 summarizes their key properties.

The normal equations, while conceptually straightforward and computationally efficient, suffer from numerical instability due to the squaring of the condition number and cannot handle rank-deficient matrices. The QR decomposition offers improved numerical stability by avoiding the formation of $X^\top X$ and provides an orthonormal basis for the column space, but still requires full column rank. The SVD is the most robust and informative method: it handles rank deficiency gracefully, provides optimal numerical stability, and reveals the complete geometric structure of the data through its singular values and singular vectors. In practice, the QR decomposition is often the method of choice for well-conditioned, full-rank problems, while the SVD is preferred when rank deficiency is suspected, when diagnostic information is needed, or when the geometric structure of the data is of independent interest.

7 Conclusion

7.1 Summary of Key Concepts

This lecture has developed a rigorous framework for understanding linear regression from first principles of linear algebra. The key results are organized as follows.

- Geometric Perspective
  - The shift from variable space ($\mathbb{R}^p$) to observation space ($\mathbb{R}^n$) transforms regression into a vector approximation problem.
  - The target vector $\mathbf{y}$ and feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p$ all reside in $\mathbb{R}^n$.
  - The set of all possible fitted values forms the column space $\mathrm{Col}(X)$.
- Projection Framework
  - The Best Approximation Theorem (Theorem 3.27): the optimal $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathrm{Col}(X)$.
  - Orthogonality condition: $\mathbf{e} \perp \mathrm{Col}(X)$, equivalently $X^\top \mathbf{e} = \mathbf{0}$.
  - This condition yields the normal equations $(X^\top X)\hat{\boldsymbol{\beta}} = X^\top \mathbf{y}$.
- Geometric-Analytical Equivalence

  - Minimizing $\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$ via calculus produces the same Normal Equations.
  - The least squares loss is the unique loss corresponding to orthogonal projection in Euclidean space.
  - Strict convexity of the loss function guarantees a unique global minimum when X has full column rank.

- Numerical Methods
  - Normal Equations: conceptually simple but numerically unstable ($\kappa(X)^2$ sensitivity).
  - QR decomposition: stable ($\kappa(X)$ sensitivity), efficient for full-rank problems.
  - SVD: most robust, handles rank deficiency, reveals complete geometric structure.

7.2 Bridge to the Workshop

In this week's workshop, you will translate this theory directly into code. You will implement the OLS estimator $\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$ from scratch using NumPy. This exercise is not merely computational; it is designed to demystify the inner workings of production-level library functions like `scikit-learn.LinearRegression`. By explicitly constructing the Gram matrix $X^T X$, computing its inverse, and solving for $\hat{\beta}$, you will see that the core of these powerful tools is simply the execution of this fundamental linear algebra equation.

7.3 Bridge to Lecture 3: Probabilistic Foundations

Our derivation today was entirely deterministic. Given a data matrix X and a target vector $\mathbf{y}$, there is one and only one OLS solution $\hat{\beta}$. However, this deterministic view provides no sense of uncertainty. How confident can we be in this estimate? If we were to collect a new dataset, would we obtain the same $\hat{\beta}$? How can we quantify the uncertainty in our predictions for new data points?

To answer these critical questions, we must move from the deterministic world of geometry to the stochastic world of probability. Next week, we will introduce a probabilistic model for the error term ϵ, assuming it is a random variable drawn from a probability distribution. We will then re-derive our result from a statistical standpoint using the principle of Maximum Likelihood Estimation (MLE). This will not only reaffirm our OLS solution under specific assumptions (namely, Gaussian errors) but will also provide the essential framework for building confidence intervals, performing hypothesis tests, and quantifying the uncertainty inherent in any data-driven model—the indispensable tools of statistical inference.

Exercises

1. **(Orthogonality Verification)** Consider the data:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 5 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 3 \\ 5 \\ 8 \\ 11 \end{pmatrix}.$$

   (a) Compute the OLS estimator $\hat{\boldsymbol{\beta}}$ using the formula $(X^\top X)^{-1} X^\top \mathbf{y}$.

   (b) Compute the residual vector $\mathbf{e} = \mathbf{y} - X\hat{\boldsymbol{\beta}}$.

   (c) Verify explicitly that $X^\top \mathbf{e} = \mathbf{0}$.

   (d) Without further computation, explain why $\mathbf{1}^\top \mathbf{e} = 0$ and $\sum_{i=1}^{n} x_i e_i = 0$, where $\mathbf{1} = (1, 1, 1, 1)^\top$ and x_i denotes the value of the second column for observation i.

2. **(Null Space and Rank Deficiency)** Consider the design matrix

$$X = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \\ 0 & 0 & 2 \end{pmatrix} \in \mathbb{R}^{4\times 3}.$$

   (a) Determine the rank of X and find a basis for $\text{Null}(X)$.

   (b) Verify that the dimension formula $\dim(\text{Col}(X)) + \dim(\text{Null}(X)) = p$ holds, where p is the number of columns.

   (c) Explain why the normal equation $(X^\top X)\boldsymbol{\beta} = X^\top \mathbf{y}$ do not have a unique solution for this matrix.

   (d) If $\mathbf{y} = (3, 6, 5, 2)^\top$, find all least squares solutions and identify the minimum-norm solution.

3. **(Right Null Space)**

   In this lecture, we studied the left null space $\text{Null}(X^\top)$ and established its relationship to the column space via Theorem 3.16. The **right null space** (or simply the **null space**) of a matrix $X \in \mathbb{R}^{n\times p}$ is defined as

$$\text{Null}(X) = \{\mathbf{v} \in \mathbb{R}^p : X\mathbf{v} = \mathbf{0}\}.$$

   This subspace consists of all vectors in the domain that are mapped to zero by X.

   Prove the following results:

   (a) Prove that the null space of X is the orthogonal complement of the row space of X:

$$\text{Null}(X) = \text{Row}(X)^\perp.$$

   Hint: Recall that $\text{Row}(X) = \text{Col}(X^\top)$.

   (b) Prove that the dimension of the null space and the rank of X satisfy

$$\dim(\text{Null}(X)) + \text{rank}(X) = p,$$

   where p is the number of columns of X.

   Hint: Use the Orthogonal Decomposition Theorem applied to $\mathbb{R}^p = \text{Row}(X) \oplus \text{Row}(X)^\perp$.

(c) Prove that for any matrix $X \in \mathbb{R}^{n \times p}$,

$$\text{Null}(X^\top X) = \text{Null}(X).$$

Use this result to explain why $X^\top X$ is invertible if and only if X has full column rank.

4. **(Generalized Pythagorean Theorem)** Prove by induction that if $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k \in V$ are mutually orthogonal vectors in an inner product space (i.e., $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ for all $i \neq j$), then

$$\left\| \sum_{i=1}^{k} \mathbf{v}_i \right\|^2 = \sum_{i=1}^{k} \|\mathbf{v}_i\|^2.$$

5. **(Properties of the Residual Maker Matrix)** Let $X \in \mathbb{R}^{n \times p}$ have linearly independent columns, and let $M = I_n - P$ where $P = X(X^\top X)^{-1} X^\top$ is the hat matrix.

   (a) Prove that M is symmetric.

   (b) Prove that M is idempotent, i.e., $M^2 = M$.

   (c) Prove that $MX = O$ (the zero matrix).

   (d) Prove that $\text{tr}(M) = n - p$.

   (e) Interpret parts (c) and (d) geometrically in terms of the column space of X.

6. **(Condition Number Analysis)** Consider the nearly collinear design matrix

$$X = \begin{pmatrix} 1 & 1.000 \\ 1 & 1.001 \\ 1 & 1.002 \end{pmatrix}.$$

   (a) Compute the Gram matrix $X^\top X$.

   (b) Find the eigenvalues of $X^\top X$ and compute the condition number $\kappa(X^\top X)$.

   (c) Explain why solving the normal equations directly would be numerically problematic for this matrix.

   (d) How does $\kappa(X^\top X)$ relate to $\kappa(X)$?

7. **(Rank and Trace of Projection Matrices)** Let $P \in \mathbb{R}^{n \times n}$ be any symmetric idempotent matrix (i.e., $P^\top = P$ and $P^2 = P$).

   (a) Prove that all eigenvalues of P are either 0 or 1. *Hint: If $P\mathbf{v} = \lambda \mathbf{v}$, apply P to both sides.*

   (b) Prove that $\text{rank}(P) = \text{tr}(P)$.

   (c) Explain how this result is consistent with Theorem 4.8.

8. **(Centering and the Intercept)** Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Define the centered variables $\tilde{x}_i = x_i - \bar{x}$ and $\tilde{y}_i = y_i - \bar{y}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

   (a) Show that the OLS estimator for the slope in the original model is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

(b) Prove that if we regress $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{x}}$ (without an intercept), the resulting coefficient equals $\hat{\beta}_1$.

(c) Prove that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

9. **(Uniqueness of the Best Approximation)** The Best Approximation Theorem states that for a finite-dimensional subspace W of an inner product space V and any $\mathbf{y} \in V$, there exists a unique vector $\hat{\mathbf{y}} \in W$ minimizing $\|\mathbf{y} - \mathbf{w}\|$ over all $\mathbf{w} \in W$. Provide a complete proof of the uniqueness assertion. *Hint: Suppose $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ both achieve the minimum, and use the parallelogram law:*

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2.$$

10. **(SVD and the Pseudoinverse)** Consider the rank-deficient matrix

$$X = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 3 \\ 7 \\ 2 \end{pmatrix}.$$

(a) Verify that X has rank 1 and find a basis for $\mathrm{Col}(X)$ and $\mathrm{Null}(X)$.

(b) Compute the SVD of X by first finding the eigenvalues and eigenvectors of $X^\top X$.

(c) Using the SVD, compute the Moore-Penrose pseudoinverse X^+.

(d) Find the minimum-norm least squares solution $\hat{\boldsymbol{\beta}} = X^+ \mathbf{y}$.

(e) Verify that among all solutions to the normal equations, $\hat{\boldsymbol{\beta}}$ has the smallest Euclidean norm.

11. **(Fundamental Subspaces)** Let $X \in \mathbb{R}^{n \times p}$ be an arbitrary matrix.

(a) Prove that $\mathrm{Row}(X) = \mathrm{Col}(X^\top)$.

(b) Prove that $\mathrm{Null}(X) = \mathrm{Row}(X)^\perp$ (where orthogonality is in $\mathbb{R}^p$).

(c) Using parts (a) and (b), derive the dimension formula: $\dim(\mathrm{Row}(X)) + \dim(\mathrm{Null}(X)) = p$.

(d) How does this relate to the SVD partition in Theorem 6.26?

12. **(Detecting Multicollinearity)** Consider the data matrix

$$X = \begin{pmatrix} 1 & 2 & 5 \\ 1 & 3 & 7 \\ 1 & 4 & 9 \\ 1 & 5 & 11 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 5 \end{pmatrix}.$$

(a) Show that the columns of X are linearly dependent by finding a non-trivial linear combination that equals the zero vector.

(b) Explain why the normal equations have infinitely many solutions.

(c) Find all solutions to the normal equations $(X^\top X)\boldsymbol{\beta} = X^\top \mathbf{y}$.

(d) Compute the projection $\hat{\mathbf{y}} = X\boldsymbol{\beta}$ and verify it is the same for all solutions found in part (c).

13. **(Orthogonality of Residuals and Fitted Values)** Let $\hat{\mathbf{y}} = P\mathbf{y}$ be the vector of fitted values and $\mathbf{e} = M\mathbf{y}$ be the residual vector, where P is the hat matrix and $M = I - P$.

(a) Prove that $\hat{\mathbf{y}}^\top \mathbf{e} = 0$, i.e., the fitted values and residuals are orthogonal.

(b) Prove the Pythagorean decomposition: $\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{e}\|^2$.

(c) Define the coefficient of determination as $R^2 = \|\hat{\mathbf{y}} - \bar{y}\mathbf{1}\|^2 / \|\mathbf{y} - \bar{y}\mathbf{1}\|^2$ where $\bar{y} = \frac{1}{n}\sum_i y_i$. Show that $0 \leq R^2 \leq 1$ when the model includes an intercept.

14. **(Equivalence of Orthogonality and First-Order Conditions)**

(a) Starting from the geometric orthogonality condition $X^\top(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{0}$, derive the normal equations.

(b) Starting from the loss function $L(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$, compute the gradient $\nabla_{\boldsymbol{\beta}} L$ and show that setting it to zero yields the same normal equations.

(c) Explain why this equivalence is not coincidental, but reflects a deep connection between orthogonal projection and least squares minimization in Euclidean space.

15. **(QR Decomposition Approach)** Let $X = QR$ be the (reduced) QR decomposition of a full column rank matrix $X \in \mathbb{R}^{n \times p}$, where $Q \in \mathbb{R}^{n \times p}$ has orthonormal columns and $R \in \mathbb{R}^{p \times p}$ is upper triangular with positive diagonal entries.

(a) Prove that $Q^\top Q = I_p$ and that $X^\top X = R^\top R$.

(b) Show that the OLS estimator satisfies $R\hat{\boldsymbol{\beta}} = Q^\top \mathbf{y}$.

(c) Prove that the hat matrix can be written as $P = QQ^\top$.

(d) Explain why solving $R\hat{\boldsymbol{\beta}} = Q^\top \mathbf{y}$ via back-substitution is numerically superior to computing $(X^\top X)^{-1} X^\top \mathbf{y}$ directly.

16. **(Leverage and Influence)** The diagonal elements h_{ii} of the hat matrix P are called *leverage* values.

(a) Prove that $0 \leq h_{ii} \leq 1$ for all i. *Hint: Use $P = P^2$ and $P = P^\top$.*

(b) Prove that $\sum_{i=1}^n h_{ii} = p + 1$ (the number of parameters including the intercept).

(c) Show that $\hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$, and interpret this as a weighted average of the observed responses.

(d) Prove that for simple linear regression with centered predictor, $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$. What does this reveal about observations with extreme predictor values?

17. **(Alternative Loss Functions and Non-Orthogonal Projections)** The manuscript establishes that minimizing $\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$ yields the orthogonal projection of $\mathbf{y}$ onto $\mathrm{Col}(X)$.

(a) Consider minimizing $\|\mathbf{y} - X\boldsymbol{\beta}\|_1 = \sum_{i=1}^n |y_i - (X\boldsymbol{\beta})_i|$ (the L^1 or LAD loss). Explain why the solution is generally *not* an orthogonal projection.

(b) Consider the weighted loss $L_W(\boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})^\top W (\mathbf{y} - X\boldsymbol{\beta})$, where $W \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix. Derive the first-order conditions and find the minimizer $\hat{\boldsymbol{\beta}}_W$.

(c) Show that the weighted least squares solution corresponds to an orthogonal projection with respect to the inner product $\langle \mathbf{u}, \mathbf{v} \rangle_W = \mathbf{u}^\top W \mathbf{v}$.

(d) What is the geometric interpretation of the weighting matrix W?

Solutions to Exercises

1. **Solution:**

   (a) We compute:

$$X^\top X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 5 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 5 \end{pmatrix} = \begin{pmatrix} 4 & 12 \\ 12 & 46 \end{pmatrix}.$$

   The determinant is $4 \cdot 46 - 12 \cdot 12 = 184 - 144 = 40$, so

$$(X^\top X)^{-1} = \frac{1}{40} \begin{pmatrix} 46 & -12 \\ -12 & 4 \end{pmatrix}.$$

   Next,

$$X^\top \mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 5 \end{pmatrix} \begin{pmatrix} 3 \\ 5 \\ 8 \\ 11 \end{pmatrix} = \begin{pmatrix} 27 \\ 98 \end{pmatrix}.$$

   Therefore,

$$\hat{\boldsymbol{\beta}} = \frac{1}{40} \begin{pmatrix} 46 & -12 \\ -12 & 4 \end{pmatrix} \begin{pmatrix} 27 \\ 98 \end{pmatrix} = \frac{1}{40} \begin{pmatrix} 1242 - 1176 \\ -324 + 392 \end{pmatrix} = \frac{1}{40} \begin{pmatrix} 66 \\ 68 \end{pmatrix} = \begin{pmatrix} 1.65 \\ 1.7 \end{pmatrix}.$$

   (b) The fitted values are:

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} 1.65 \\ 1.7 \end{pmatrix} = \begin{pmatrix} 3.35 \\ 5.05 \\ 8.45 \\ 10.15 \end{pmatrix}.$$

   The residuals are:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} 3 - 3.35 \\ 5 - 5.05 \\ 8 - 8.45 \\ 11 - 10.15 \end{pmatrix} = \begin{pmatrix} -0.35 \\ -0.05 \\ -0.45 \\ 0.85 \end{pmatrix}.$$

   (c) We verify:

$$X^\top \mathbf{e} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 5 \end{pmatrix} \begin{pmatrix} -0.35 \\ -0.05 \\ -0.45 \\ 0.85 \end{pmatrix} = \begin{pmatrix} -0.35 - 0.05 - 0.45 + 0.85 \\ -0.35 - 0.10 - 1.80 + 4.25 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

   Note: Due to rounding, we get approximately zero. With exact arithmetic using $\hat{\beta}_0 = 33/20$ and $\hat{\beta}_1 = 17/10$, we obtain exactly $X^\top \mathbf{e} = \mathbf{0}$.

   (d) The condition $X^\top \mathbf{e} = \mathbf{0}$ means that each column of X is orthogonal to $\mathbf{e}$. The first column of X is $\mathbf{1}$, so $\mathbf{1}^\top \mathbf{e} = 0$, which is equivalent to $\sum_i e_i = 0$. The second column contains the x_i values, so its orthogonality to $\mathbf{e}$ gives $\sum_i x_i e_i = 0$.

2. **Solution:**

(a) We perform row reduction on X to determine its rank:

$$X = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \\ 0 & 0 & 2 \end{pmatrix} \xrightarrow{R_2 - 2R_1} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 1 & 2 & 3 \\ 0 & 0 & 2 \end{pmatrix} \xrightarrow{R_3 - R_1} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 2 \end{pmatrix} \xrightarrow{R_4 - R_3} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

The row echelon form has two pivots (in columns 1 and 3), so $\operatorname{rank}(X) = 2$. The second column is a free variable. To find $\operatorname{Null}(X)$, we solve $X\mathbf{v} = \mathbf{0}$. Setting $v_2 = t$ (free parameter), the system gives $v_3 = 0$ from the third row, and $v_1 + 2v_2 + v_3 = 0$ from the first row, yielding $v_1 = -2t$. Thus,

$$\operatorname{Null}(X) = \operatorname{span}\left\{ \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} \right\}.$$

A basis for $\operatorname{Null}(X)$ is $\{(-2, 1, 0)^\top\}$.

(b) We verify the Rank-Nullity Theorem. The number of columns is $p = 3$. We have:

$$\dim(\operatorname{Col}(X)) + \dim(\operatorname{Null}(X)) = \operatorname{rank}(X) + \operatorname{nullity}(X) = 2 + 1 = 3 = p. \quad \checkmark$$

(c) Since $\operatorname{Null}(X) \neq \{\mathbf{0}\}$, the columns of X are linearly dependent. By Theorem 4.1, this implies $\operatorname{Null}(X^\top X) = \operatorname{Null}(X) \neq \{\mathbf{0}\}$, and hence the Gram matrix $X^\top X$ is singular by Theorem 4.2. Consequently, the Normal Equations cannot be solved by multiplying by $(X^\top X)^{-1}$.

Geometrically, multiple coefficient vectors $\boldsymbol{\beta}$ produce the same fitted value $\hat{\mathbf{y}} = X\boldsymbol{\beta}$. Specifically, if $\hat{\boldsymbol{\beta}}$ is any solution, then $\hat{\boldsymbol{\beta}} + \mathbf{v}$ is also a solution for any $\mathbf{v} \in \operatorname{Null}(X)$, since $X(\hat{\boldsymbol{\beta}} + \mathbf{v}) = X\hat{\boldsymbol{\beta}} + X\mathbf{v} = X\hat{\boldsymbol{\beta}}$.

(d) We first find one particular least squares solution. Since the second column of X equals twice the first column, we can eliminate redundancy by working with a reduced matrix. Setting $\beta_2 = 0$ and solving with the remaining columns:

$$X_{\text{reduced}} = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 1 & 3 \\ 0 & 2 \end{pmatrix}, \quad X_{\text{reduced}}^\top X_{\text{reduced}} = \begin{pmatrix} 6 & 10 \\ 10 & 18 \end{pmatrix}, \quad X_{\text{reduced}}^\top \mathbf{y} = \begin{pmatrix} 20 \\ 34 \end{pmatrix}.$$

Solving $(X_{\text{reduced}}^\top X_{\text{reduced}})\boldsymbol{\gamma} = X_{\text{reduced}}^\top \mathbf{y}$:

$$(X_{\text{reduced}}^\top X_{\text{reduced}})^{-1} = \frac{1}{8}\begin{pmatrix} 18 & -10 \\ -10 & 6 \end{pmatrix}, \quad \boldsymbol{\gamma} = \frac{1}{8}\begin{pmatrix} 18 & -10 \\ -10 & 6 \end{pmatrix}\begin{pmatrix} 20 \\ 34 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 0.5 \end{pmatrix}.$$

Thus one particular solution is $\hat{\boldsymbol{\beta}}_0 = (2.5, 0, 0.5)^\top$. The general least squares solution is

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 + t\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2.5 - 2t \\ t \\ 0.5 \end{pmatrix}, \quad t \in \mathbb{R}.$$

To find the minimum-norm solution, we minimize $\|\hat{\boldsymbol{\beta}}\|_2^2 = (2.5 - 2t)^2 + t^2 + 0.25$ over t. Taking the derivative and setting it to zero:

$$\frac{d}{dt}\left[(2.5 - 2t)^2 + t^2\right] = -4(2.5 - 2t) + 2t = 10t - 10 = 0 \implies t = 1.$$

The minimum-norm solution is therefore

$$\hat{\boldsymbol{\beta}}^{+} = \begin{pmatrix} 2.5 - 2 \\ 1 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1 \\ 0.5 \end{pmatrix}.$$

This is the solution that would be obtained via the Moore-Penrose pseudoinverse: $\hat{\boldsymbol{\beta}}^{+} = X^{+}\mathbf{y}$.

3. **The Right Null Space and Its Properties:**

   (a) **Orthogonal Complement of the Row Space:**
   We prove that $\text{Null}(X) = \text{Row}(X)^{\perp}$ by establishing inclusion in both directions.
   First, suppose $\mathbf{v} \in \text{Null}(X)$, so that $X\mathbf{v} = \mathbf{0}$. Let $\mathbf{w} \in \text{Row}(X)$ be arbitrary. Since $\text{Row}(X) = \text{Col}(X^{\top})$, there exists $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that $\mathbf{w} = X^{\top}\boldsymbol{\alpha}$. The inner product of $\mathbf{v}$ with $\mathbf{w}$ is

   $$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^{\top}\mathbf{w} = \mathbf{v}^{\top}X^{\top}\boldsymbol{\alpha} = (X\mathbf{v})^{\top}\boldsymbol{\alpha} = \mathbf{0}^{\top}\boldsymbol{\alpha} = 0.$$

   Since $\mathbf{w}$ was arbitrary, $\mathbf{v} \perp \text{Row}(X)$, and hence $\mathbf{v} \in \text{Row}(X)^{\perp}$.
   Conversely, suppose $\mathbf{v} \in \text{Row}(X)^{\perp}$, so that $\mathbf{v}$ is orthogonal to every vector in $\text{Row}(X)$. In particular, $\mathbf{v}$ is orthogonal to each row of X. Let $\mathbf{r}_1^{\top}, \mathbf{r}_2^{\top}, \ldots, \mathbf{r}_n^{\top}$ denote the rows of X, so that $\mathbf{r}_i \in \text{Row}(X)$ for each i. Then

   $$\langle \mathbf{v}, \mathbf{r}_i \rangle = \mathbf{r}_i^{\top}\mathbf{v} = 0 \text{ for all } i.$$

   The product $X\mathbf{v}$ is a vector whose i-th component is $\mathbf{r}_i^{\top}\mathbf{v} = 0$. Therefore, $X\mathbf{v} = \mathbf{0}$, which means $\mathbf{v} \in \text{Null}(X)$.
   Having established both inclusions, we conclude that $\text{Null}(X) = \text{Row}(X)^{\perp}$.

   (b) **Rank-Nullity Theorem:**
   By part (a), $\text{Null}(X) = \text{Row}(X)^{\perp}$. Applying the Orthogonal Decomposition Theorem to the vector space $\mathbb{R}^p$, we have

   $$\mathbb{R}^p = \text{Row}(X) \oplus \text{Row}(X)^{\perp} = \text{Row}(X) \oplus \text{Null}(X).$$

   For a direct sum decomposition, the dimensions satisfy

   $$\dim(\mathbb{R}^p) = \dim(\text{Row}(X)) + \dim(\text{Null}(X)).$$

   Since $\dim(\mathbb{R}^p) = p$ and $\dim(\text{Row}(X)) = \text{rank}(X)$ (the row rank equals the column rank), we obtain

   $$p = \text{rank}(X) + \dim(\text{Null}(X)),$$

   which is equivalent to $\dim(\text{Null}(X)) + \text{rank}(X) = p$.

   (c) **Characterization via the Gram Matrix:**
   We prove $\text{Null}(X^{\top}X) = \text{Null}(X)$ by showing inclusion in both directions.
   For $\text{Null}(X) \subseteq \text{Null}(X^{\top}X)$: Let $\mathbf{v} \in \text{Null}(X)$, so $X\mathbf{v} = \mathbf{0}$. Then

   $$(X^{\top}X)\mathbf{v} = X^{\top}(X\mathbf{v}) = X^{\top}\mathbf{0} = \mathbf{0},$$

   hence $\mathbf{v} \in \text{Null}(X^{\top}X)$.

For $\text{Null}(X^\top X) \subseteq \text{Null}(X)$: Let $\mathbf{v} \in \text{Null}(X^\top X)$, so $(X^\top X)\mathbf{v} = \mathbf{0}$. Left-multiplying by $\mathbf{v}^\top$ gives $\mathbf{v}^\top X^\top X\mathbf{v} = 0$. Recognizing this as $\|X\mathbf{v}\|_2^2 = 0$, positive definiteness of the norm implies $X\mathbf{v} = \mathbf{0}$, hence $\mathbf{v} \in \text{Null}(X)$.

Therefore, $\text{Null}(X^\top X) = \text{Null}(X)$.

For the invertibility statement: The matrix $X^\top X \in \mathbb{R}^{p\times p}$ is invertible if and only if $\text{Null}(X^\top X) = \{\mathbf{0}\}$. By the equality just proven, this holds if and only if $\text{Null}(X) = \{\mathbf{0}\}$. By Theorem 4.1, $\text{Null}(X) = \{\mathbf{0}\}$ if and only if the columns of X are linearly independent, i.e., X has full column rank. Thus, $X^\top X$ is invertible if and only if X has full column rank.

4. **Solution:**

Base case ($k = 2$): This is the standard Pythagorean theorem proven in the manuscript. For orthogonal vectors $\mathbf{v}_1, \mathbf{v}_2$:

$$\|\mathbf{v}_1 + \mathbf{v}_2\|^2 = \langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_1 + \mathbf{v}_2\rangle = \|\mathbf{v}_1\|^2 + 2\langle \mathbf{v}_1, \mathbf{v}_2\rangle + \|\mathbf{v}_2\|^2 = \|\mathbf{v}_1\|^2 + \|\mathbf{v}_2\|^2.$$

Inductive step: Assume the result holds for k mutually orthogonal vectors. Consider $k+1$ mutually orthogonal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}$. Let $\mathbf{w} = \sum_{i=1}^k \mathbf{v}_i$. We claim that $\mathbf{w} \perp \mathbf{v}_{k+1}$:

$$\langle \mathbf{w}, \mathbf{v}_{k+1}\rangle = \left\langle \sum_{i=1}^k \mathbf{v}_i, \mathbf{v}_{k+1} \right\rangle = \sum_{i=1}^k \langle \mathbf{v}_i, \mathbf{v}_{k+1}\rangle = 0,$$

since each $\mathbf{v}_i$ is orthogonal to $\mathbf{v}_{k+1}$ by the mutual orthogonality assumption.

Applying the base case to $\mathbf{w}$ and $\mathbf{v}_{k+1}$:

$$\left\| \sum_{i=1}^{k+1} \mathbf{v}_i \right\|^2 = \|\mathbf{w} + \mathbf{v}_{k+1}\|^2 = \|\mathbf{w}\|^2 + \|\mathbf{v}_{k+1}\|^2.$$

By the inductive hypothesis, $\|\mathbf{w}\|^2 = \sum_{i=1}^k \|\mathbf{v}_i\|^2$. Therefore:

$$\left\| \sum_{i=1}^{k+1} \mathbf{v}_i \right\|^2 = \sum_{i=1}^k \|\mathbf{v}_i\|^2 + \|\mathbf{v}_{k+1}\|^2 = \sum_{i=1}^{k+1} \|\mathbf{v}_i\|^2.$$

By induction, the result holds for all $k \geq 2$. □

5. **Solution:**

(a) Since P is symmetric ($P^\top = P$), we have:

$$M^\top = (I_n - P)^\top = I_n^\top - P^\top = I_n - P = M.$$

(b) Using the idempotence of P ($P^2 = P$):

$$\begin{aligned} M^2 &= (I_n - P)(I_n - P) = I_n - P - P + P^2 \\ &= I_n - P - P + P = I_n - P = M. \end{aligned}$$

(c) We compute:

$$MX = (I_n - P)X = X - PX = X - X(X^\top X)^{-1}X^\top X = X - X \cdot I_p = X - X = O.$$

(d) Using the linearity and cyclic property of trace:

$$\text{tr}(M) = \text{tr}(I_n - P) = \text{tr}(I_n) - \text{tr}(P) = n - (p+1).$$

If the design matrix has p features plus an intercept column, then $\text{tr}(M) = n - (p+1)$.

(e) Part (c) shows that M annihilates every column of X, meaning M projects onto the orthogonal complement of $\text{Col}(X)$. Part (d) confirms this: the dimension of $\text{Col}(X)^\perp$ is $n - \text{rank}(X) = n - p$ (assuming full column rank), which equals $\text{tr}(M)$ since for projection matrices, trace equals rank.

6. **Solution:**

(a) We compute:

$$X^\top X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1.001 & 1.002 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1.001 \\ 1 & 1.002 \end{pmatrix} = \begin{pmatrix} 3 & 3.003 \\ 3.003 & 3.006005 \end{pmatrix}.$$

(b) The characteristic polynomial is:

$$\det(X^\top X - \lambda I) = (3-\lambda)(3.006005 - \lambda) - 3.003^2 = \lambda^2 - 6.006005\lambda + 0.000006.$$

Using the quadratic formula:

$$\lambda = \frac{6.006005 \pm \sqrt{36.072... - 0.000024}}{2} \approx \frac{6.006005 \pm 6.005995}{2}.$$

So $\lambda_1 \approx 6.006$ and $\lambda_2 \approx 0.000005 = 5 \times 10^{-6}$.

The condition number is:

$$\kappa(X^\top X) = \frac{\lambda_{\max}}{\lambda_{\min}} \approx \frac{6.006}{5 \times 10^{-6}} \approx 1.2 \times 10^6.$$

(c) With $\kappa(X^\top X) \approx 10^6$, small perturbations in the data (of order machine epsilon $\epsilon \approx 10^{-16}$) can cause errors in $\hat{\boldsymbol{\beta}}$ of order $\kappa \cdot \epsilon \approx 10^{-10}$. While this may seem acceptable, the effective loss of precision means only about 10 significant digits remain reliable. For problems with larger condition numbers, catastrophic cancellation can occur.

(d) If σ_1, σ_2 are the singular values of X, then $\lambda_i = \sigma_i^2$ are the eigenvalues of $X^\top X$. Thus:

$$\kappa(X^\top X) = \frac{\sigma_1^2}{\sigma_2^2} = \left(\frac{\sigma_1}{\sigma_2}\right)^2 = \kappa(X)^2.$$

The condition number squares when forming the normal equations, explaining why methods avoiding this (QR, SVD) are preferred.

7. **Solution:**

(a) Let λ be an eigenvalue of P with eigenvector $\mathbf{v} \neq \mathbf{0}$, so $P\mathbf{v} = \lambda\mathbf{v}$. Applying P to both sides and using idempotence:

$$P^2\mathbf{v} = P(\lambda\mathbf{v}) = \lambda P\mathbf{v} = \lambda^2\mathbf{v}.$$

But $P^2 = P$, so $P^2\mathbf{v} = P\mathbf{v} = \lambda\mathbf{v}$. Therefore $\lambda\mathbf{v} = \lambda^2\mathbf{v}$, which gives $(\lambda - \lambda^2)\mathbf{v} = \mathbf{0}$. Since $\mathbf{v} \neq \mathbf{0}$, we have $\lambda(1-\lambda) = 0$, so $\lambda = 0$ or $\lambda = 1$.

(b) Since P is symmetric, it is diagonalizable with an orthonormal basis of eigenvectors. Let r be the number of eigenvalues equal to 1 (counting multiplicity), and $n-r$ be the number equal to 0. Then:

$$\operatorname{tr}(P) = \sum_{i=1}^{n} \lambda_i = r \cdot 1 + (n-r) \cdot 0 = r.$$

The rank of P equals the dimension of its range, which equals the number of nonzero eigenvalues (for diagonalizable matrices), which is r. Therefore $\operatorname{rank}(P) = r = \operatorname{tr}(P)$.

(c) For the hat matrix $P = X(X^\top X)^{-1}X^\top$, we proved $\operatorname{tr}(P) = p+1$ (the number of columns of the design matrix). By part (b), this equals $\operatorname{rank}(P)$. Since P projects onto $\operatorname{Col}(X)$, and $\operatorname{rank}(P) = \dim(\operatorname{Col}(X)) = p+1$, the results are consistent.

8. **Solution:**

(a) For simple linear regression with design matrix having columns $\mathbf{1}$ and $\mathbf{x} = (x_1, \ldots, x_n)^\top$, the normal equations give:

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}.$$

Using $\bar{x} = \frac{1}{n}\sum x_i$ and $\bar{y} = \frac{1}{n}\sum y_i$, and the identities:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y},$$
$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2,$$

we can verify that:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.$$

(b) Regressing $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{x}}$ without intercept, the normal equation is $(\tilde{\mathbf{x}}^\top \tilde{\mathbf{x}})\gamma = \tilde{\mathbf{x}}^\top \tilde{\mathbf{y}}$. Thus:

$$\gamma = \frac{\tilde{\mathbf{x}}^\top \tilde{\mathbf{y}}}{\tilde{\mathbf{x}}^\top \tilde{\mathbf{x}}} = \frac{\sum \tilde{x}_i \tilde{y}_i}{\sum \tilde{x}_i^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \hat{\beta}_1.$$

(c) From the normal equations, the first equation (corresponding to the intercept column $\mathbf{1}$) gives:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i.$$

Dividing by n: $\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}$, so $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

9. **Solution:**

Suppose $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2 \in W$ both minimize $\|\mathbf{y} - \mathbf{w}\|$ over $\mathbf{w} \in W$. Let $d = \|\mathbf{y} - \hat{\mathbf{y}}_1\| = \|\mathbf{y} - \hat{\mathbf{y}}_2\|$ be the minimum distance.

Since W is a subspace, the midpoint $\mathbf{m} = \frac{1}{2}(\hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2) \in W$. By the triangle inequality, $\|\mathbf{y} - \mathbf{m}\| \geq d$ (since d is the minimum).

Apply the parallelogram law with $\mathbf{u} = \mathbf{y} - \hat{\mathbf{y}}_1$ and $\mathbf{v} = \mathbf{y} - \hat{\mathbf{y}}_2$:

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2 = 2d^2 + 2d^2 = 4d^2.$$

Note that $\mathbf{u}+\mathbf{v} = 2\mathbf{y} - \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 = 2(\mathbf{y}-\mathbf{m})$, so $\|\mathbf{u}+\mathbf{v}\|^2 = 4\|\mathbf{y}-\mathbf{m}\|^2 \geq 4d^2$.

Also, $\mathbf{u}-\mathbf{v} = \hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1$.

From the parallelogram law:

$$4\|\mathbf{y}-\mathbf{m}\|^2 + \|\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1\|^2 = 4d^2.$$

Since $\|\mathbf{y}-\mathbf{m}\|^2 \geq d^2$, we have $4d^2 + \|\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1\|^2 \leq 4d^2$, which implies $\|\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1\|^2 \leq 0$.

Since norms are non-negative, $\|\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1\| = 0$, hence $\hat{\mathbf{y}}_1 = \hat{\mathbf{y}}_2$. □

10. **Solution:**

(a) The second column is twice the first: $(2,4,2)^\top = 2(1,2,1)^\top$. Thus $\text{rank}(X) = 1$.
Basis for $\text{Col}(X)$: $\{(1,2,1)^\top\}$.
For $\text{Null}(X)$: we solve $X\boldsymbol{\beta} = \mathbf{0}$, i.e., $\beta_1(1,2,1)^\top + \beta_2(2,4,2)^\top = \mathbf{0}$. This gives $\beta_1 + 2\beta_2 = 0$, so $\text{Null}(X) = \text{span}\{(2,-1)^\top\}$.

(b) We compute $X^\top X$ and its eigendecomposition:

$$X^\top X = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 6 & 12 \\ 12 & 24 \end{pmatrix}.$$

Eigenvalues: $\det(X^\top X - \lambda I) = \lambda^2 - 30\lambda = \lambda(\lambda - 30) = 0$, so $\lambda_1 = 30$, $\lambda_2 = 0$.
Singular values: $\sigma_1 = \sqrt{30}$, $\sigma_2 = 0$.
Eigenvector for $\lambda_1 = 30$: $(1,2)^\top/\sqrt{5}$. So $V_1 = (1,2)^\top/\sqrt{5}$.
The left singular vector: $U_1 = XV_1/\sigma_1 = \frac{1}{\sqrt{30}}(1,2,1)^\top \cdot \sqrt{5} = (1,2,1)^\top/\sqrt{6}$.

(c) The pseudoinverse is:

$$X^+ = V_1\sigma_1^{-1}U_1^\top = \frac{1}{\sqrt{5}}\begin{pmatrix}1\\2\end{pmatrix} \cdot \frac{1}{\sqrt{30}} \cdot \frac{1}{\sqrt{6}} \begin{pmatrix}1 & 2 & 1\end{pmatrix} = \frac{1}{30}\begin{pmatrix}1 & 2 & 1\\ 2 & 4 & 2\end{pmatrix}.$$

(d)

$$\hat{\boldsymbol{\beta}} = X^+\mathbf{y} = \frac{1}{30}\begin{pmatrix}1 & 2 & 1\\ 2 & 4 & 2\end{pmatrix}\begin{pmatrix}3\\7\\2\end{pmatrix} = \frac{1}{30}\begin{pmatrix}3+14+2\\6+28+4\end{pmatrix} = \frac{1}{30}\begin{pmatrix}19\\38\end{pmatrix} = \begin{pmatrix}19/30\\19/15\end{pmatrix}.$$

(e) The general solution to the normal equations is $\hat{\boldsymbol{\beta}} + t(2,-1)^\top$ for $t \in \mathbb{R}$. The norm squared is:

$$\left\| \begin{pmatrix} 19/30 + 2t \\ 19/15 - t \end{pmatrix} \right\|^2 = (19/30 + 2t)^2 + (19/15 - t)^2.$$

Taking the derivative with respect to t and setting to zero:

$$2(19/30+2t)(2) + 2(19/15 - t)(-1) = 0 \implies 4(19/30 + 2t) = 19/15 - t \implies t = 0.$$

Thus $\hat{\boldsymbol{\beta}} = X^+\mathbf{y}$ is indeed the minimum-norm solution.

11. **Solution:**

(a) By definition, $\text{Row}(X) = \text{span}\{\text{rows of } X\}$. The rows of X are the columns of $X^\top$, so $\text{Row}(X) = \text{span}\{\text{columns of } X^\top\} = \text{Col}(X^\top)$.

(b) We show $\text{Null}(X) \subseteq \text{Row}(X)^\perp$: Let $\mathbf{v} \in \text{Null}(X)$, so $X\mathbf{v} = \mathbf{0}$. For any row $\mathbf{r}_i^\top$ of X, the i-th component of $X\mathbf{v}$ is $\mathbf{r}_i^\top \mathbf{v} = 0$. Thus $\mathbf{v}$ is orthogonal to every row, hence $\mathbf{v} \in \text{Row}(X)^\perp$.

For the reverse inclusion: Let $\mathbf{v} \in \text{Row}(X)^\perp$. Then $\mathbf{v}$ is orthogonal to each row of X, meaning $\mathbf{r}_i^\top \mathbf{v} = 0$ for all i. But these are precisely the components of $X\mathbf{v}$, so $X\mathbf{v} = \mathbf{0}$ and $\mathbf{v} \in \text{Null}(X)$.

(c) From part (b) and the orthogonal decomposition theorem:

$$\mathbb{R}^p = \text{Row}(X) \oplus \text{Row}(X)^\perp = \text{Row}(X) \oplus \text{Null}(X).$$

Therefore $p = \dim(\text{Row}(X)) + \dim(\text{Null}(X))$.

(d) In the SVD partition, V_1 spans $\text{Row}(X)$ (dimension r) and V_2 spans $\text{Null}(X)$ (dimension $p - r$). The orthogonality of V's columns ensures $\text{Row}(X) \perp \text{Null}(X)$, confirming part (b).

12. **Solution:**

(a) Observe that $(1, 2, 5)^\top + 2(0, 1, 2)^\top = (1, 4, 9)^\top$, i.e., $\mathbf{x}_1 + 2\mathbf{x}_2 = \mathbf{x}_3$ where $\mathbf{x}_j$ denotes the j-th column. Thus $(1, 2, -1)^\top$ is in the null space: $(1)\mathbf{x}_1 + (2)\mathbf{x}_2 + (-1)\mathbf{x}_3 = \mathbf{0}$.

(b) Since the columns are linearly dependent, $X^\top X$ is singular and hence not invertible. The normal equations $(X^\top X)\boldsymbol{\beta} = X^\top \mathbf{y}$ have infinitely many solutions (a solution exists since $X^\top \mathbf{y} \in \text{Col}(X^\top X)$).

(c) We compute $X^\top X$ and $X^\top \mathbf{y}$:

$$X^\top X = \begin{pmatrix} 4 & 14 & 32 \\ 14 & 54 & 122 \\ 32 & 122 & 276 \end{pmatrix}, \quad X^\top \mathbf{y} = \begin{pmatrix} 11 \\ 42 \\ 95 \end{pmatrix}.$$

One can verify the third column of $X^\top X$ equals the first plus twice the second, consistent with the dependency. A particular solution is found by setting $\beta_3 = 0$ and solving the reduced system for β_0, β_1. The general solution is:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0^* \\ \beta_1^* \\ 0 \end{pmatrix} + t \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \quad t \in \mathbb{R},$$

where (β_0^*, β_1^*) solves the reduced normal equations.

(d) The projection $\hat{\mathbf{y}} = X\boldsymbol{\beta}$ is independent of the choice of $\boldsymbol{\beta}$ among solutions because:

$$X(\boldsymbol{\beta} + t\mathbf{n}) = X\boldsymbol{\beta} + tX\mathbf{n} = X\boldsymbol{\beta} + t\mathbf{0} = X\boldsymbol{\beta},$$

where $\mathbf{n} = (1, 2, -1)^\top \in \text{Null}(X)$.

13. **Solution:**

(a) Using $\hat{\mathbf{y}} = P\mathbf{y}$ and $\mathbf{e} = M\mathbf{y} = (I - P)\mathbf{y}$:

$$\hat{\mathbf{y}}^\top \mathbf{e} = (P\mathbf{y})^\top (I - P)\mathbf{y} = \mathbf{y}^\top P^\top (I - P)\mathbf{y} = \mathbf{y}^\top P(I - P)\mathbf{y},$$

using symmetry $P^\top = P$. Now:

$$P(I - P) = P - P^2 = P - P = O,$$

using idempotence. Therefore $\hat{\mathbf{y}}^\top \mathbf{e} = 0$.

(b) Since $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$ and $\hat{\mathbf{y}} \perp \mathbf{e}$ (from part a), the Pythagorean theorem gives:

$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}} + \mathbf{e}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{e}\|^2.$$

(c) With an intercept, $\mathbf{1} \in \text{Col}(X)$, so $P\mathbf{1} = \mathbf{1}$ and thus $\bar{\hat{y}} = \bar{y}$. Writing $\mathbf{y} - \bar{y}\mathbf{1} = (\hat{\mathbf{y}} - \bar{y}\mathbf{1}) + \mathbf{e}$, and noting these are orthogonal (since $\mathbf{e} \perp \text{Col}(X) \ni (\hat{\mathbf{y}} - \bar{y}\mathbf{1})$):

$$\|\mathbf{y} - \bar{y}\mathbf{1}\|^2 = \|\hat{\mathbf{y}} - \bar{y}\mathbf{1}\|^2 + \|\mathbf{e}\|^2.$$

Thus:

$$R^2 = \frac{\|\hat{\mathbf{y}} - \bar{y}\mathbf{1}\|^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|^2} = 1 - \frac{\|\mathbf{e}\|^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|^2}.$$

Since $\|\mathbf{e}\|^2 \geq 0$ and $\|\mathbf{e}\|^2 \leq \|\mathbf{y} - \bar{y}\mathbf{1}\|^2$, we have $0 \leq R^2 \leq 1$.

14. **Solution:**

(a) The orthogonality condition states $X^\top(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{0}$. Expanding:

$$X^\top\mathbf{y} - X^\top X\boldsymbol{\beta} = \mathbf{0} \implies X^\top X\boldsymbol{\beta} = X^\top\mathbf{y}.$$

These are the normal equations.

(b) Expanding the loss function:

$$L(\boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})^\top(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top X^\top\mathbf{y} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}.$$

Taking the gradient (using $\nabla_{\boldsymbol{\beta}}(\boldsymbol{\beta}^\top A\boldsymbol{\beta}) = 2A\boldsymbol{\beta}$ for symmetric A):

$$\nabla_{\boldsymbol{\beta}} L = -2X^\top\mathbf{y} + 2X^\top X\boldsymbol{\beta}.$$

Setting this to zero: $X^\top X\boldsymbol{\beta} = X^\top\mathbf{y}$, the same normal equations.

(c) The equivalence reflects that in Euclidean space (with the standard inner product), orthogonal projection minimizes squared distance. The gradient ∇L points in the direction of steepest ascent of L. Setting $\nabla L = 0$ finds stationary points. For the Euclidean norm, the gradient being zero is equivalent to the residual being orthogonal to the constraint space. This deep connection only holds for L^2 (Euclidean) norms; other norms yield different optimality conditions.

15. **Solution:**

(a) Since Q has orthonormal columns: $Q^\top Q = I_p$.
For $X^\top X$: $X^\top X = (QR)^\top(QR) = R^\top Q^\top QR = R^\top I_p R = R^\top R$.

(b) Starting from the normal equations:

$$X^\top X\hat{\boldsymbol{\beta}} = X^\top\mathbf{y} \implies R^\top R\hat{\boldsymbol{\beta}} = R^\top Q^\top\mathbf{y}.$$

Since R is invertible (positive diagonal entries), so is $R^\top$. Left-multiplying by $(R^\top)^{-1}$:

$$R\hat{\boldsymbol{\beta}} = Q^\top\mathbf{y}.$$

(c) We verify:

$$P = X(X^\top X)^{-1}X^\top = QR(R^\top R)^{-1}R^\top Q^\top = QRR^{-1}(R^\top)^{-1}R^\top Q^\top = QI_pQ^\top = QQ^\top.$$

(d) The equation $R\hat{\boldsymbol{\beta}} = Q^\top \mathbf{y}$ is an upper triangular system solvable by back-substitution in $O(p^2)$ operations. This avoids:

- Forming $X^\top X$ (which squares the condition number)
- Computing $(X^\top X)^{-1}$ (expensive and unstable)

The QR approach has condition number $\kappa(R) = \kappa(X)$, versus $\kappa(X)^2$ for normal equations.

16. **Solution:**

(a) From $P^2 = P$ and $P^\top = P$, we have $h_{ii} = P_{ii} = (P^2)_{ii} = \sum_{j=1}^n P_{ij}P_{ji} = \sum_{j=1}^n P_{ij}^2 \geq P_{ii}^2 = h_{ii}^2$.
Thus $h_{ii} \geq h_{ii}^2$, which gives $h_{ii}(1 - h_{ii}) \geq 0$. Combined with $h_{ii} = \sum_j P_{ij}^2 \geq 0$, we get $0 \leq h_{ii} \leq 1$.

(b) $\sum_{i=1}^n h_{ii} = \operatorname{tr}(P) = p + 1$ by Theorem 4.8.

(c) From $\hat{\mathbf{y}} = P\mathbf{y}$, the i-th component is:

$$\hat{y}_i = \sum_{j=1}^n P_{ij}y_j = P_{ii}y_i + \sum_{j \neq i} P_{ij}y_j = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j.$$

This shows $\hat{y}_i$ is a weighted combination of all responses, with weight h_{ii} on the i-th observation itself.

(d) For simple linear regression with centered predictor, the design matrix is $X = (\mathbf{1}, \tilde{\mathbf{x}})$ where $\tilde{x}_i = x_i - \bar{x}$. The hat matrix has diagonal elements:

$$h_{ii} = \frac{1}{n} + \frac{\tilde{x}_i^2}{\sum_j \tilde{x}_j^2} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}.$$

The minimum leverage is $1/n$ (when $x_i = \bar{x}$). Observations with extreme predictor values have higher leverage, meaning they exert more influence on the fitted line. High-leverage points can drastically affect regression coefficients if they are also outliers in the response.

17. **Solution:**

(a) The L^1 loss $\|\mathbf{y} - X\boldsymbol{\beta}\|_1 = \sum_i |y_i - (X\boldsymbol{\beta})_i|$ is not derived from an inner product. The L^1 geometry uses the "taxicab" metric, where the unit ball is a cross-polytope rather than a sphere. The minimizer of the L^1 loss produces a vector $\hat{\mathbf{y}}$ in $\operatorname{Col}(X)$, but the residual $\mathbf{y} - \hat{\mathbf{y}}$ is generally *not* orthogonal to $\operatorname{Col}(X)$ in the Euclidean sense. Instead, the optimality conditions involve subdifferentials and median-like properties.

(b) Setting the gradient to zero:

$$\nabla_{\boldsymbol{\beta}} L_W = -2X^\top W(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{0}.$$

This gives $X^\top W X\boldsymbol{\beta} = X^\top W\mathbf{y}$, so:

$$\hat{\boldsymbol{\beta}}_W = (X^\top W X)^{-1} X^\top W\mathbf{y}.$$

(c) Define the weighted inner product $\langle \mathbf{u}, \mathbf{v} \rangle_W = \mathbf{u}^\top W \mathbf{v}$. This is a valid inner product since W is symmetric positive definite. The induced norm is $\|\mathbf{u}\|_W = \sqrt{\mathbf{u}^\top W \mathbf{u}}$.

The first-order condition $X^\top W(\mathbf{y} - X\hat{\boldsymbol{\beta}}_W) = \mathbf{0}$ says that for each column $\mathbf{x}_j$ of X:

$$\langle \mathbf{x}_j, \mathbf{y} - X\hat{\boldsymbol{\beta}}_W \rangle_W = \mathbf{x}_j^\top W(\mathbf{y} - X\hat{\boldsymbol{\beta}}_W) = 0.$$

Thus the residual is orthogonal to $\mathrm{Col}(X)$ *in the W-inner product*, making $X\hat{\boldsymbol{\beta}}_W$ the orthogonal projection of $\mathbf{y}$ onto $\mathrm{Col}(X)$ with respect to $\langle \cdot, \cdot \rangle_W$.

(d) The matrix W changes the geometry of $\mathbb{R}^n$ by defining a new notion of distance. If $W = \mathrm{diag}(w_1, \ldots, w_n)$ is diagonal, then $\|\mathbf{u}\|_W^2 = \sum_i w_i u_i^2$, so observations with larger w_i contribute more to the loss. Geometrically, W stretches or compresses coordinates, transforming the Euclidean sphere into an ellipsoid. The weighted least squares solution finds the point in $\mathrm{Col}(X)$ closest to $\mathbf{y}$ in this deformed space.