# Data Science for Mathematicians

## Lesson 2: Linear Regression from a Geometric Perspective

Department of Mathematics and Computer Science

December 18, 2025

# Outline

1. From Data Clouds to Best-Fit Subspaces
2. The Algebraic Formulation
3. The Orthogonal Projection Principle
4. Derivation of the Estimator
5. The Hat Matrix and Projection
6. Minimization via Calculus
7. Convex Optimization Perspective
8. Numerical Methods and Stability
9. Summary and Preview

# Two Views of the Data Matrix

Consider a dataset with $n$ observations and $p$ features: $X \in \mathbb{R}^{n \times p}$

**Variable Space View (Scatter Plot):**

- $n$ row vectors $\{\mathbf{x}_i^T\}_{i=1}^n$ in $\mathbb{R}^p$
- Each vector $=$ one observation
- Intuitive for $p = 2$ or $p = 3$

**Observation Space View (Today's Focus):**

- $p$ column vectors $\{\mathbf{x}_j\}_{j=1}^p$ in $\mathbb{R}^n$
- Each vector $=$ all observations for one feature
- Algebraically powerful!

# The Observation Space View

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

**Key insight:** This transforms regression from:

*"Fitting a hyperplane to points"* $\longrightarrow$ *"Vector approximation"*

**Why is this powerful?**

- Treat entire feature as a single vector
- Apply linear algebra: subspaces, orthogonality, projections
- Handle the whole dataset simultaneously

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

**Given:**

- Feature vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\} \subset \mathbb{R}^n$
- Target vector $\mathbf{y} \in \mathbb{R}^n$

**Goal:** Find the best linear approximation:

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

**Geometric interpretation:** $\hat{\mathbf{y}}$ is constructed by scaling and adding feature vectors.

# The Column Space

**Definition: Column Space**

The **column space** of $X$ is the set of all linear combinations of its columns:

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

**Key properties:**

- $\mathrm{Col}(X)$ is a subspace of $\mathbb{R}^n$
- $\dim(\mathrm{Col}(X)) = \mathrm{rank}(X) \leq \min(n, p)$
- Every possible fitted value $\hat{\mathbf{y}} = X\boldsymbol{\beta}$ lives in $\mathrm{Col}(X)$

# Example: Column Space

## Example

Consider the data matrix:

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

The column space is:

$$\mathrm{Col}(X) = \left\{ \beta_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \beta_2 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_1 + \beta_2 \end{pmatrix} : \beta_1, \beta_2 \in \mathbb{R} \right\}$$

**Geometrically:** A 2D plane through the origin in $\mathbb{R}^3$

# Example: Linear Dependence (Multicollinearity)

## Example

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

Although $X$ has 2 columns:

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

**Geometrically:** Only a 1D line (not a plane!)

This is **multicollinearity**: $\mathrm{rank}(X) = 1 < 2$

# The Geometric Goal of Linear Regression

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

From linear algebra, we know:

- The closest vector in a subspace is the **orthogonal projection**

**Therefore:**

$$\boxed{\hat{\mathbf{y}} = \mathrm{proj}_{\mathrm{Col}(X)}\mathbf{y}}$$

# The Linear Model in Matrix Form

For observation $i$:

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

**Design matrix** (with intercept column of 1s):

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

where:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

# Why No Exact Solution Exists

**Overdetermined system:** $n \gg p + 1$ (more equations than unknowns)

**The problem:**

- $\mathbf{y} = X\boldsymbol{\beta}$ has a solution $\Leftrightarrow$ $\mathbf{y} \in \mathrm{Col}(X)$
- Due to noise/error, $\mathbf{y}$ almost never lies in $\mathrm{Col}(X)$
- The target vector "sticks out" of the column space

**Geometric meaning of error:**

$$\boldsymbol{\epsilon} = \mathbf{y} - X\boldsymbol{\beta}$$

is the displacement from $\mathrm{Col}(X)$ to $\mathbf{y}$.

**Our goal:** Find $\hat{\boldsymbol{\beta}}$ that makes this displacement *as short as possible*.

# Definition: Inner Product

## Definition: Inner Product

An **inner product** on vector space $V$ is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ satisfying:

1. *Symmetry*: $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$
2. *Linearity*: $\langle \alpha\mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \alpha\langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$
3. *Positive definiteness*: $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$, with equality iff $\mathbf{u} = \mathbf{0}$

## Example: Euclidean Inner Product

In $\mathbb{R}^n$:

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

# Definition: Induced Norm and Distance

## Definition: Induced Norm

The **norm** induced by an inner product is:

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

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

## Example

In $\mathbb{R}^n$, this gives the **Euclidean norm**:

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

For $\mathbf{v} = (3, -4, 0)^\top$: $\|\mathbf{v}\|_2 = \sqrt{9 + 16 + 0} = 5$

# Definition: Orthogonality

## Definition: Orthogonality

Two vectors are **orthogonal**, written $\mathbf{u} \perp \mathbf{v}$, if:

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

A vector $\mathbf{v}$ is **orthogonal to a subspace** $W$ if $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ for all $\mathbf{w} \in W$.

## Example

In $\mathbb{R}^3$: $\mathbf{u} = (1, 2, -1)^\top$ and $\mathbf{v} = (3, 0, 3)^\top$

$$\langle \mathbf{u}, \mathbf{v} \rangle = 1 \cdot 3 + 2 \cdot 0 + (-1) \cdot 3 = 3 - 3 = 0 \quad \checkmark$$

# Definition: Orthogonal Complement

## Definition: Orthogonal Complement

The **orthogonal complement** of subspace $W$ is:

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

## Example

Let $W = \mathrm{span}\{(1,0,1)^\top\}$ in $\mathbb{R}^3$.
Find $W^\perp$: Need $v_1 + v_3 = 0$, so $v_3 = -v_1$.

$$W^\perp = \{(v_1, v_2, -v_1)^\top : v_1, v_2 \in \mathbb{R}\} = \mathrm{span}\{(1,0,-1)^\top, (0,1,0)^\top\}$$

**Note:** $\dim(W) + \dim(W^\perp) = 1 + 2 = 3 = \dim(\mathbb{R}^3)$

# Definition: Left Null Space

## Definition: Left Null Space

The **left null space** of $X \in \mathbb{R}^{n \times p}$ is:

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

## Theorem

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

*The left null space equals the orthogonal complement of the column space.*

**Why this matters:** If $\mathbf{e} \perp \mathrm{Col}(X)$, then $X^\top \mathbf{e} = \mathbf{0}$

# Definition: Direct Sum

## Definition: Direct Sum

$V = U \oplus W$ (direct sum) if:

1. *Spanning*: Every $\mathbf{v} \in V$ can be written as $\mathbf{v} = \mathbf{u} + \mathbf{w}$
2. *Trivial intersection*: $U \cap W = \{\mathbf{0}\}$

When both hold, the decomposition is *unique*.

## Example

In $\mathbb{R}^2$: Let $U = x$-axis, $W = y$-axis.
Any $(a, b)^\top = (a, 0)^\top + (0, b)^\top$ uniquely.
$\therefore \mathbb{R}^2 = U \oplus W$

# Theorem: Orthogonal Decomposition

### Theorem: Orthogonal Decomposition

Let $W$ be a finite-dimensional subspace of inner product space $V$. Then:

$$V = W \oplus W^{\perp}$$

**Meaning:** Every vector $\mathbf{v}$ can be *uniquely* written as:

$$\mathbf{v} = \underbrace{\mathbf{w}}_{\in W} + \underbrace{\mathbf{w}^{\perp}}_{\in W^{\perp}}$$

with $\mathbf{w} \perp \mathbf{w}^{\perp}$.

# Definition: Orthogonal Projection

## Definition: Orthogonal Projection

The **orthogonal projection** of $\mathbf{y}$ onto subspace $W$, denoted $\mathrm{proj}_W(\mathbf{y})$, is the unique vector in $W$ such that:

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

**From the Orthogonal Decomposition:**

$$\mathbf{y} = \underbrace{\mathrm{proj}_W(\mathbf{y})}_{\text{component in } W} + \underbrace{(\mathbf{y} - \mathrm{proj}_W(\mathbf{y}))}_{\text{component in } W^\perp}$$

# Theorem: Pythagorean Theorem

**Theorem: Pythagorean Theorem**

If $\mathbf{u} \perp \mathbf{v}$, then:

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

**Proof:**

$$\begin{aligned} \|\mathbf{u}+\mathbf{v}\|^2 &= \langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + 2\underbrace{\langle \mathbf{u}, \mathbf{v} \rangle}_{=0} + \langle \mathbf{v}, \mathbf{v} \rangle \\ &= \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 \end{aligned}$$

# Definition: Orthogonal Projection Matrix

## Definition: Orthogonal Projection Matrix

A matrix $P \in \mathbb{R}^{n \times n}$ is an **orthogonal projection matrix** if:

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

**Interpretation:**

- **Symmetry:** Respects inner product structure
- **Idempotence:** Projecting twice $=$ projecting once

# Example: Projection onto a Line

## Example

Project onto $W = \mathrm{span}\{(1,1)^\top\}$ in $\mathbb{R}^2$.
Unit vector: $\mathbf{u} = \frac{1}{\sqrt{2}}(1,1)^\top$

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

**Verify:**

- $P^\top = P$ ✓
- $P^2 = \frac{1}{4}\begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} = P$ ✓

For $\mathbf{v} = (3,1)^\top$: $P\mathbf{v} = (2,2)^\top$
Residual: $(3,1)^\top - (2,2)^\top = (1,-1)^\top \perp (1,1)^\top$ ✓

# Theorem: Best Approximation

**Theorem: Best Approximation Theorem**

Let $W$ be a subspace of inner product space $V$, and $\mathbf{y} \in V$.
The orthogonal projection $\mathrm{proj}_W(\mathbf{y})$ is the **unique** vector in $W$ closest to $\mathbf{y}$:

$$\|\mathbf{y} - \mathrm{proj}_W(\mathbf{y})\| < \|\mathbf{y} - \mathbf{w}\| \quad \text{for all } \mathbf{w} \in W, \mathbf{w} \neq \mathrm{proj}_W(\mathbf{y})$$

**This is why orthogonal projection solves linear regression!**

# Proof of Best Approximation Theorem

Let $\hat{\mathbf{y}} = \mathrm{proj}_W(\mathbf{y})$. For any $\mathbf{w} \in W$:

$$\mathbf{y} - \mathbf{w} = \underbrace{(\mathbf{y} - \hat{\mathbf{y}})}_{\in W^\perp} + \underbrace{(\hat{\mathbf{y}} - \mathbf{w})}_{\in W}$$

These two vectors are orthogonal! By Pythagorean theorem:

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

**Equality** holds iff $\|\hat{\mathbf{y}} - \mathbf{w}\|^2 = 0$ iff $\mathbf{w} = \hat{\mathbf{y}}$.

# The Orthogonality Condition

**Applying Best Approximation to Regression:**

- Vector space: $V = \mathbb{R}^n$
- Subspace: $W = \mathrm{Col}(X)$
- Best approximation: $\hat{\mathbf{y}} = \mathrm{proj}_{\mathrm{Col}(X)}(\mathbf{y})$

The **residual** $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ must satisfy:

$$\mathbf{e} \perp \mathrm{Col}(X) \quad \Leftrightarrow \quad \mathbf{e} \in \mathrm{Null}(X^\top)$$

**Equivalently:** $\mathbf{e}$ is orthogonal to every column of $X$:

$$\mathbf{x}_j^\top \mathbf{e} = 0 \quad \text{for } j = 0, 1, \ldots, p$$

# The Master Equation of Orthogonality

Stacking all orthogonality conditions:

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

$$\boxed{X^\top \mathbf{e} = \mathbf{0}}$$

**Interpretation:** Residuals are uncorrelated with every predictor.
**Special case:** $\mathbf{x}_0 = \mathbf{1}_n$ implies $\sum_i e_i = 0$ (residuals sum to zero).

# Deriving the Normal Equations

Since $\hat{\mathbf{y}} \in \mathrm{Col}(X)$, we can write $\hat{\mathbf{y}} = X\hat{\beta}$.

Substituting $\mathbf{e} = \mathbf{y} - X\hat{\beta}$ into $X^\top \mathbf{e} = \mathbf{0}$:

$$
\begin{aligned}
X^\top(\mathbf{y} - X\hat{\beta}) &= \mathbf{0} \\
X^\top \mathbf{y} - X^\top X\hat{\beta} &= \mathbf{0}
\end{aligned}
$$

**The Normal Equations:**

$$\boxed{(X^\top X)\hat{\beta} = X^\top \mathbf{y}}$$

**Why "normal"?** The residual must be *normal* (perpendicular) to $\mathrm{Col}(X)$.

# The Gram Matrix

**Definition: Gram Matrix**

$G = X^\top X \in \mathbb{R}^{(p+1)\times(p+1)}$ is the **Gram matrix**.
Its $(i, j)$-entry is $G_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i^\top \mathbf{x}_j$

**Properties:**

- Symmetric: $(X^\top X)^\top = X^\top X$
- Diagonal entries: $G_{ii} = \|\mathbf{x}_i\|^2$ (squared norms)
- Off-diagonal: inner products between features

**Theorem**

$X^\top X$ is invertible $\Leftrightarrow$ columns of $X$ are linearly independent.

If columns of $X$ are linearly independent, $X^\top X$ is invertible.

Solving the Normal Equations:

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

**The OLS Estimator:**

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

This is the **Ordinary Least Squares (OLS)** estimator—the cornerstone of linear regression!

# Example: Simple Linear Regression

## Example

Data: $(1,2),(2,3),(3,5)$

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

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

**Step 2:** $X^\top \mathbf{y} = \begin{pmatrix} 10 \\ 23 \end{pmatrix}$

**Step 3:** $\hat{\beta} = \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}\begin{pmatrix} 10 \\ 23 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 3/2 \end{pmatrix}$

**Fitted model:** $\hat{y} = \frac{1}{3} + \frac{3}{2}x$

# The Hat Matrix (Projection Matrix)

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

**Definition: Hat Matrix**

The **hat matrix** (or projection matrix) is:

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

It "puts a hat on $\mathbf{y}$": $\hat{\mathbf{y}} = P\mathbf{y}$

**Theorem**

The hat matrix $P$ is symmetric ($P^\top = P$) and idempotent ($P^2 = P$).

Therefore, $P$ is an orthogonal projection matrix onto $\mathrm{Col}(X)$.

# Theorem: Trace of the Hat Matrix

**Theorem**

For a design matrix $X \in \mathbb{R}^{n\times(p+1)}$ with linearly independent columns:

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

**Proof (using cyclic property of trace):**

$$\begin{aligned}
\mathrm{tr}(P) &= \mathrm{tr}(X(X^\top X)^{-1}X^\top) \\
&= \mathrm{tr}(X^\top X(X^\top X)^{-1}) \\
&= \mathrm{tr}(I_{p+1}) \\
&= p + 1
\end{aligned}$$

**Interpretation:** The trace equals the number of parameters.

# The Residual Maker Matrix

## Definition: Residual Maker Matrix

$$M = I_n - P$$

**Properties:**

- Extracts residuals: $\mathbf{e} = M\mathbf{y}$
- Projects onto $\mathrm{Col}(X)^\perp = \mathrm{Null}(X^\top)$
- Also symmetric and idempotent
- $PM = MP = O$ (orthogonality of subspaces)

**Orthogonal decomposition:**

$$\mathbf{y} = \underbrace{P\mathbf{y}}_{\hat{\mathbf{y}}} + \underbrace{M\mathbf{y}}_{\mathbf{e}} \quad \text{with } \hat{\mathbf{y}} \perp \mathbf{e}$$

# Example: Hat Matrix Computation

## Example

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

$$P = \frac{1}{6}\begin{pmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix}$$

**Diagonal entries** $h_{ii}$ are called **leverages**:

- Measure influence of observation $i$ on its own fitted value
- Here: $h_{11} = h_{33} = 5/6$, $h_{22} = 2/6 = 1/3$
- Extreme observations have higher leverage

**Verify:** $\mathrm{tr}(P) = \frac{5+2+5}{6} = 2 = p + 1$ ✓

# The Least Squares Objective

**Analytical formulation:** Minimize the Sum of Squared Residuals (SSR):

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

**Key observation:**

*Minimizing squared error = Finding shortest residual vector*

This connects the analytical and geometric perspectives!

# Gradient Calculation

Expand the loss function:

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

Using matrix calculus rules:

- $\nabla_{\boldsymbol{\beta}}(\mathbf{a}^\top \boldsymbol{\beta}) = \mathbf{a}$
- $\nabla_{\boldsymbol{\beta}}(\boldsymbol{\beta}^\top A \boldsymbol{\beta}) = 2A\boldsymbol{\beta}$ (for symmetric $A$)

$$\nabla_{\boldsymbol{\beta}} L = -2X^\top \mathbf{y} + 2X^\top X \boldsymbol{\beta}$$

Setting $\nabla L = \mathbf{0}$:

$$X^\top X \boldsymbol{\beta} = X^\top \mathbf{y} \quad \leftarrow \textbf{The Normal Equations again!}$$

# The Unity of Geometry and Analysis

**Two paths, same destination:**

**Geometric:**

- Orthogonal projection
- Residual $\perp$ column space
- $X^\top \mathbf{e} = \mathbf{0}$

**Analytical:**

- Minimize squared error
- Set gradient to zero
- $\nabla L = \mathbf{0}$

Both yield: $(X^\top X)\hat{\beta} = X^\top \mathbf{y}$

**Deep insight:** Least squares is not arbitrary—it's the unique loss function corresponding to orthogonal projection in Euclidean space!

# Definition: Convex Function

### Definition: Convex Function

A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if for all $\mathbf{x}, \mathbf{z}$ and $\lambda \in [0, 1]$:

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

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

**Geometric interpretation:** The line segment between any two points on the graph lies above (or on) the graph.

## Example: $f(x) = x^2$ is Strictly Convex

### Example

For $f(x) = x^2$, with $x \neq z$ and $\lambda \in (0, 1)$:

$$\lambda f(x) + (1-\lambda)f(z) - f(\lambda x + (1-\lambda)z) = \lambda(1-\lambda)(x-z)^2 > 0$$

Since $\lambda(1-\lambda) > 0$ and $(x-z)^2 > 0$.

**More generally:** $f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$ is:

- Convex if $A$ is positive semi-definite
- Strictly convex if $A$ is positive definite

# Theorem: Local Minima of Convex Functions

**Theorem**

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Then:

1. Every local minimum is also a **global minimum**
2. If $f$ is strictly convex, there is **at most one** global minimum

**Theorem: Hessian Characterization**

If $f$ is twice differentiable:

- $f$ convex $\Leftrightarrow$ Hessian $H(\mathbf{x}) \succeq 0$ (positive semi-definite) $\forall \mathbf{x}$
- $f$ strictly convex $\Leftarrow$ $H(\mathbf{x}) \succ 0$ (positive definite) $\forall \mathbf{x}$

# Theorem: Convexity of the Least Squares Loss

## Theorem

The least squares loss $L(\beta) = \|\mathbf{y} - X\beta\|_2^2$ is:

- Always convex
- Strictly convex $\Leftrightarrow$ columns of $X$ are linearly independent

**Proof:** The Hessian is $H = 2X^\top X$.
For any $\mathbf{v}$:

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

Strictly positive (for $\mathbf{v} \neq \mathbf{0}$) iff $X\mathbf{v} \neq \mathbf{0}$ iff columns are linearly independent.

**Conclusion:** When $X$ has full column rank, the OLS estimator is the **unique global minimum**.

# Definition: Condition Number

## Definition: Condition Number

The **condition number** of an invertible matrix $A$ is:

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

Using the spectral norm:

$$\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

**Interpretation:**

- $\kappa(A)$ measures sensitivity of $A\mathbf{x} = \mathbf{b}$ to perturbations
- $\kappa(A) \approx 1$: well-conditioned
- $\kappa(A) \gg 1$: ill-conditioned (errors amplified!)

# Example: Ill-Conditioned Matrix

## Example

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1.0001 \end{pmatrix}, \quad \kappa(A) \approx 20000$$

Solve $A\mathbf{x} = \mathbf{b}$ with $\mathbf{b} = (2, 2.0001)^\top$. Solution: $\mathbf{x} = (1, 1)^\top$

Perturb to $\tilde{\mathbf{b}} = (2, 2.0002)^\top$ (0.005% change)
New solution: $\tilde{\mathbf{x}} = (0, 2)^\top$ (100% change!)

**Lesson:** Tiny input errors $\to$ huge output errors

# The Problem with Normal Equations

**Critical issue:** Forming $X^\top X$ squares the condition number!

$$\kappa(X^\top X) = \kappa(X)^2$$

**Disaster scenario:**

- $X$ has near-collinear columns: $\kappa(X) \approx 10^7$
- $X^\top X$: $\kappa(X^\top X) \approx 10^{14}$
- Double precision: $\sim 16$ digits
- Could lose $\sim 14$ digits of precision!

**Conclusion:** Never directly compute $(X^\top X)^{-1}$ in practice!

# Definition: QR Decomposition

## Definition: QR Decomposition

For $X \in \mathbb{R}^{n \times m}$ with $n \geq m$:

$$X = QR$$

where:

- $Q \in \mathbb{R}^{n \times m}$ has orthonormal columns ($Q^\top Q = I_m$)
- $R \in \mathbb{R}^{m \times m}$ is upper triangular

## Theorem

Every matrix with $n \geq m$ has a QR decomposition. If $X$ has full column rank, $R$ can have positive diagonal entries (unique).

# QR Approach to Least Squares

Substitute $X = QR$ into the objective:

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

Minimized when:

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

**Theorem**

$\kappa(R) = \kappa(X)$ (no squaring!)

**Advantages:**

- Condition number not squared
- $R$ is triangular: solve by back-substitution ($O(m^2)$)
- No matrix inversion needed!

# Definition: Singular Value Decomposition (SVD)

## Definition: SVD

For any $X \in \mathbb{R}^{n \times m}$:

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

where:

- $U \in \mathbb{R}^{n \times n}$ is orthogonal (left singular vectors)
- $V \in \mathbb{R}^{m \times m}$ is orthogonal (right singular vectors)
- $\Sigma \in \mathbb{R}^{n \times m}$ is diagonal with $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ (singular values)

## Theorem

Every matrix has an SVD. Singular values are unique; they equal the square roots of the eigenvalues of $X^\top X$.

# Definition: Moore-Penrose Pseudoinverse

## Definition: Pseudoinverse

For $X = U\Sigma V^\top$, the **Moore-Penrose pseudoinverse** is:

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

where $\Sigma^+$ is obtained by transposing $\Sigma$ and replacing each nonzero $\sigma_i$ by $1/\sigma_i$.

## Example

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

# Theorem: Moore-Penrose Conditions

## Theorem: Moore-Penrose Conditions

$X^+$ is the unique matrix satisfying:

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

## Theorem

For full column rank $X$:

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

# SVD Solution to Least Squares

**Theorem**

The vector $\hat{\beta} = X^{+}\mathbf{y}$ is the least squares solution with minimum norm.

**Advantages of SVD:**

1. **Numerical stability:** Condition number $= \kappa(X)$, not $\kappa(X)^2$
2. **Rank deficiency:** Handles singular $X^\top X$ gracefully
3. **Diagnostics:** Singular values reveal near-collinearity

**Formula:**

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

# Comparison of Methods

| | **Normal Equations** | **QR** | **SVD** |
|---|---|---|---|
| Key equation | $(X^\top X)^{-1}X^\top \mathbf{y}$ | $R^{-1}Q^\top \mathbf{y}$ | $V\Sigma^+U^\top \mathbf{y}$ |
| Condition # | $\kappa(X)^2$ | $\kappa(X)$ | $\kappa(X)$ |
| Rank deficient | No | No | **Yes** |
| Geometric insight | Minimal | Moderate | **Full** |

**Practical recommendations:**

- QR: Default for well-conditioned, full-rank problems
- SVD: When rank deficiency suspected or diagnostics needed
- Never: Direct normal equations computation

# Theorem: SVD and the Four Fundamental Subspaces

**Theorem**

Let $X = U\Sigma V^\top$ with $r = \mathrm{rank}(X)$. Partition $U = (U_1|U_2)$ and $V = (V_1|V_2)$ where $U_1, V_1$ have $r$ columns. Then:

1. Columns of $V_1$: orthonormal basis for $\mathrm{Row}(X)$
2. Columns of $V_2$: orthonormal basis for $\mathrm{Null}(X)$
3. Columns of $U_1$: orthonormal basis for $\mathrm{Col}(X)$
4. Columns of $U_2$: orthonormal basis for $\mathrm{Null}(X^\top)$

**The SVD reveals the complete geometric structure of $X$!**

# Key Takeaways

**Geometric Perspective:**

- Shift from variable space ($\mathbb{R}^p$) to observation space ($\mathbb{R}^n$)
- Linear regression = projection onto $\mathrm{Col}(X)$
- Residual must be orthogonal to column space

**Analytical Perspective:**

- Minimize squared error $\|\mathbf{y} - X\boldsymbol{\beta}\|^2$
- Both approaches yield the Normal Equations

**Key Formulas:**

$$
\begin{aligned}
\text{OLS Estimator:} \quad & \hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y} \\
\text{Hat Matrix:} \quad & P = X(X^\top X)^{-1} X^\top \\
\text{Fitted Values:} \quad & \hat{\mathbf{y}} = P\mathbf{y}
\end{aligned}
$$