# Linear Models: Probability and Linear Algebra Review

Robert Gulotty

University of Chicago

February 25, 2026

# Two Ways an Empirical Project Can Fail

## 1) Identification Failure

- The causal parameter (e.g. $\tau$) is not uniquely determined by the observable data.
- Multiple causal stories are observationally equivalent.

**Implication:**

*Your research question cannot be answered with these data.*

## 2) Estimation / Inference Failure

- The causal parameter *is* determined by the data under your assumptions.
- But the estimator targets the wrong object or uncertainty is mismeasured.

**Implication:**

*The question is answerable — but your numerical answer or reported certainty may be wrong.*

# Example: Identification Is Fine, Estimation Is Not

- Suppose we run a block-randomized experiment,
- Treatment is assigned at the state level,
- Outcomes are observed at the county level.

**Identification:**

- Difference in means across counties identifies the ATE.

**Estimation Problem:**

- Outcomes are correlated within states.
- Naive OLS treats counties as independent.

**Consequence:**

- The coefficient is consistent.
- Standard errors can be severely understated (10x or more!).

*The research question is answerable — but the reported certainty may be false.*

# Goals for Today

- This is a condensed review of probability theory and linear algebra.
- We focus on the concepts that connect directly to econometric practice:
  1. Expectation, variance, and why they matter for prediction.
  2. Joint, marginal, and conditional distributions.
  3. The Conditional Expectation Function (CEF) as the target of regression.
  4. Matrix algebra and the geometry of projection.
  5. Positive definiteness, eigenvalues, and degrees of freedom.

# Expectation

## Definition

*Expected Value: define the expected value of $Y$ as,*

$$\begin{aligned}\mu = \mathbb{E}[Y] &= \sum_{j=1} \tau_j P[Y = \tau_j] && \textit{when } Y \textit{ takes on discrete values } \tau \\ = \mathbb{E}[Y] &= \int_{-\infty}^{\infty} y f(y) dy && \textit{when } Y \textit{ is continuous}\end{aligned}$$

*For all values of $x$ with $p(x)$ greater than zero, take the sum/integral of values times the probability weights.*

Unified notation (Riemann-Stieltjes): $\mathbb{E}[X] = \int_{-\infty}^{\infty} x \, dF(x)$

# Key Properties of Expectation

The fact that expected values are sums/integrals gives us the following properties, for random variable X and Y, constant a.

$$
\begin{aligned}
\mathbb{E}[X+Y] &= \mathbb{E}[X] + \mathbb{E}[Y] \\
\mathbb{E}[a] &= a \\
\mathbb{E}[aX+b] &= a\mathbb{E}[X] + b \\
\mathbb{E}[\mathbb{E}[X]] &= \mathbb{E}[X] \\
\mathbb{E}[XY] &\neq \mathbb{E}[X] \times \mathbb{E}[Y] \quad \text{(in general)}
\end{aligned}
$$

# Expectation Minimizes Mean Squared Error

If we want to predict $y$ with no other information, and our prediction is $\mu$, minimize:

$$\begin{aligned} M &= \mathbb{E}[(y-\mu)^2] \\ &= \mathbb{E}[y^2] - 2\mu\mathbb{E}[y] + \mu^2 \end{aligned}$$

Using calculus to minimize:

$$\begin{aligned} \frac{d}{d\mu}M &= -2\mathbb{E}[y] + 2\mu = 0 \\ \mu^* &= \mathbb{E}[y] \end{aligned}$$

This is a special case of the fact that the *conditional expectation function* minimizes mean-square prediction error.

# Variance

## Definition

*The variance of a random variable $X$, $var(X)$, is*

$$\begin{aligned} Var(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \end{aligned}$$

- Standard deviation: $\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$, population variance: $\sigma^2$.

## Corollary

$Var(aX + b) = a^2 Var(X)$

# Sample Variance: Two Equivalent Forms

$$
\begin{aligned}
\widehat{\mathrm{Var}}(X) &= \frac{1}{N}\sum_i (x_i - \bar{x})^2 \\
&= \frac{1}{N}\sum_i (x_i - \bar{x})(x_i - \bar{x}) \\
&= \frac{1}{N}\sum_i \left[(x_i - \bar{x})x_i - (x_i - \bar{x})\bar{x}\right] \\
&= \frac{1}{N}\sum_i (x_i - \bar{x})x_i \;-\; \bar{x}\cdot \underbrace{\frac{1}{N}\sum_i (x_i - \bar{x})}_{=0} \\
&= \frac{1}{N}\sum_i (x_i - \bar{x})x_i
\end{aligned}
$$

The key step: $\frac{1}{N}\sum_i (x_i - \bar{x}) = \bar{x} - \bar{x} = 0$.
This identity—that deviations from the mean sum to zero—will reappear when we derive OLS.

# Covariance

## Definition

*The* **covariance** *of two random variables X and Y is*

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Key properties:

- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(aX + b, cY + d) = ac\,\text{Cov}(X, Y)$
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y)$
- If $X \perp Y$, then $\text{Cov}(X, Y) = 0$. The converse is false in general.

# Variance of Linear Combinations (Matrix Form)

- Scalar: $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\,\text{Cov}(X, Y)$
- For a random vector $\boldsymbol{X} = (X_1, \ldots, X_k)'$, define the **variance-covariance matrix**:

$$\text{Var}(\boldsymbol{X}) = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])'] = \boldsymbol{\Sigma}$$

  This is a $k \times k$ symmetric, positive semi-definite matrix.
- For any fixed matrix $\boldsymbol{A}$ $(m \times k)$ and vector $\boldsymbol{b}$ $(m \times 1)$:

$$\text{Var}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A}\,\text{Var}(\boldsymbol{X})\,\boldsymbol{A}' = \boldsymbol{A}\Sigma\boldsymbol{A}'$$

- This formula appears throughout the course:
  - Variance of $\hat{\beta}$: $\text{Var}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}\,\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$
  - Sandwich formula, GLS, robust standard errors—all follow this pattern.

# Moments and Regularity Conditions

- $\mathbb{E}[X^r] = \int_{-\infty}^{\infty} x^r \, dF(x)$ is the $r$th *moment* of $X$.
- For some distributions, the expectation, the variance, or "higher" moments may not be finite.
- When $\mathbb{E}[X^r] = \infty$, the $r$th moment does not exist.
- Examples:
  - Fat tails distributions (e.g. Pareto distributions) often have no finite variance
  - Ratios: if $X$, $Y$ are independent standard normal, $Z = X/Y$ has no finite expectation.
- Many econometric results require finite second (or fourth) moments to offer probabilistic guarantees.
- Feel free to assume all moments exist, but I'll try to be careful.

# The Normal Distribution

## Definition

*$X \sim Normal(\mu, \sigma^2)$ has density*

$$f(x) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Key properties:

- If $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b,\ a^2\sigma^2)$.
- If $X_1 \perp X_2$ and both normal, $X_1 + X_2 \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2)$.
- If $X_1$ and $X_2$ are *jointly* normal and uncorrelated, then they are independent.

# Moments of the Normal Distribution

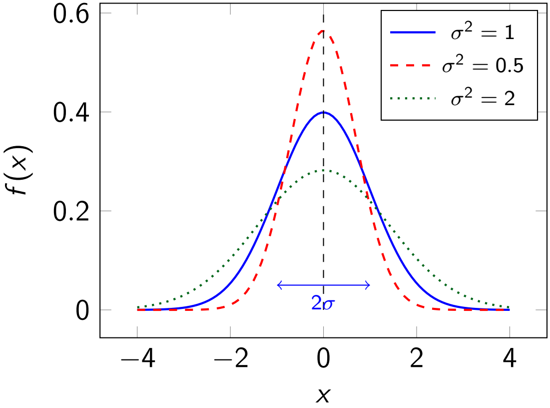
Let $X \sim N(\mu, \sigma^2)$.

## Raw Moments

- First moment: $\mathbb{E}[X] = \mu$
- Second moment: $\mathbb{E}[X^2] = \sigma^2 + \mu^2$

## Central Moments

- Variance: $\mathbb{E}[(X-\mu)^2] = \sigma^2$
- Third central moment: $\mathbb{E}[(X-\mu)^3] = 0 \quad \text{(symmetric)}$
- Fourth central moment: $\mathbb{E}[(X-\mu)^4] = 3\sigma^4$



# Moment Conditions

A parameter can be defined by an expectation it must satisfy.

**Example 1: Mean**

$$\mathbb{E}[X - \mu] = 0.$$

The true value of $\mu$ is the one that makes this expectation zero.

**Example 2: Linear Regression**

$$\mathbb{E}\left[X(Y - X'\beta)\right] = 0.$$

The true $\beta$ is the one that makes the regressors, X, uncorrelated with the error $(Y - X'\beta)$.

# General Form

Many models can be written as:

$$\mathbb{E}[g(W, \theta)] = 0.$$

- $W$ = observable data.
- $\theta$ = parameter.
- $g(\cdot)$ encodes the economic or statistical restrictions.

Identification requires that these conditions determine a unique $\theta$.

# Joint Distributions

### Definition

*The joint distribution function of* $(X, Y)$ *is* $F(x, y) = P[X \leq x,\ Y \leq y]$.

### Definition

*The joint density is* $f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$.

If our data is continuous, densities and distributions live in 3 (or higher) dimensions.

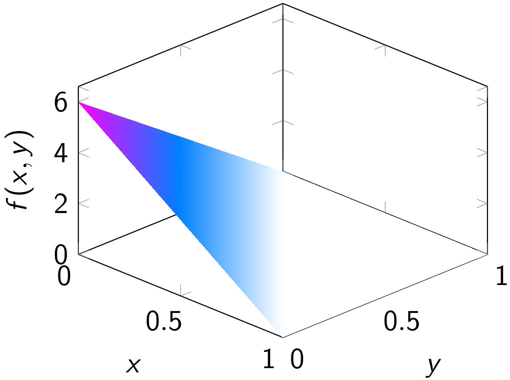
# Example: A Joint Density

Consider two parties (A, B) mobilizing voters, with $X + Y \leq 1$:

$$f(x, y) = 6(1 - x - y), \quad x, y \geq 0$$

- Density is highest at $(0, 0)$ and decreases as either party mobilizes more.
- The 6 ensures $\int\int f(x, y)\, dx\, dy = 1$:

$$\int_0^1 \int_0^{1-x} 6(1 - x - y)\, dy\, dx = 1$$



# Marginal and Conditional Densities

## Definition

*The* **marginal density** *of $X$ is*

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

## Definition

*The* **conditional density** *of $Y$ given $X = x$ is*

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}$$

*for any $x$ such that $f_X(x) > 0$.*

These definitions are the bridge from joint distributions to regression.

# Example: Marginal and Conditional from $f(x,y) = 6(1-x-y)$

**Marginal density of $X$** fix x, integrate out $y$:

$$f_X(x) = \int_0^{1-x} 6(1-x-y)\,dy = 6\left[(1-x)y - \frac{y^2}{2}\right]_0^{1-x} = 3(1-x)^2$$

**Conditional density of $Y|X=x$:**

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{6(1-x-y)}{3(1-x)^2} = \frac{2(1-x-y)}{(1-x)^2}$$

Note: for each fixed $x$, this is a valid density in $y$ on $[0, 1-x]$.

# Conditional Expectation

$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y \, f_{Y|X}(y|x) \, dy = \frac{\int_{-\infty}^{\infty} y \, f(y, x) \, dy}{\int_{-\infty}^{\infty} f(y, x) \, dy}$$

*The average value of $Y$ given that $X$ equals the specific value $x$.*

# Conditional Expectation Function (CEF)

- **CEF**:

$$\mathbb{E}[Y|X = x] = m(x)$$

- $Y$ is the dependent variable, $X = (X_1, \ldots, X_k)'$ are the independent variables.
- $m(x) = \mathbb{E}[Y|X = x]$ is the value of a function at the real value $x$.
- $m(X) = \mathbb{E}[Y|X]$ is a function of a random variable, so is itself a random variable.
- We will show that the CEF is the best predictor of $Y$ given $X$ in the mean-square error sense.
- In most applications we use a *linear approximation* to the CEF, then make inferences about the joint distribution.

# Example: Computing the CEF from $f(x,y) = 6(1-x-y)$

Apply the definition using our conditional density:

$$
\begin{aligned}
\mathbb{E}[Y|X=x] &= \int_0^{1-x} y\, f_{Y|X}(y|x)\, dy \\
&= \int_0^{1-x} y \cdot \frac{2(1-x-y)}{(1-x)^2}\, dy \\
&= \frac{2}{(1-x)^2} \int_0^{1-x} \left[(1-x)y - y^2\right] dy \\
&= \frac{2}{(1-x)^2} \left[\frac{(1-x)y^2}{2} - \frac{y^3}{3}\right]_0^{1-x} \\
&= \frac{2}{(1-x)^2} \left[\frac{(1-x)^3}{2} - \frac{(1-x)^3}{3}\right] = \frac{2}{(1-x)^2} \cdot \frac{(1-x)^3}{6} \\
&= \frac{1-x}{3}
\end{aligned}
$$

# Law of Iterated Expectations

## Theorem

$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[Y|X]] &= \int_{-\infty}^{\infty} \mathbb{E}[Y|X=x]\, f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, dy\, f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, \frac{f(y,x)}{f_X(x)} f_X(x)\, dy\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, f(y,x)\, dy\, dx \\
&= \int_{-\infty}^{\infty} y\, f_Y(y)\, dy = \mathbb{E}[Y]
\end{aligned}
$$

# Law of Total Variance

**Theorem**

$Var[Y] = \mathbb{E}[Var[Y|X]] + Var[\mathbb{E}[Y|X]]$

- $\text{Var}[\mathbb{E}[Y|X]]$: variance of the CEF — the "explained" variance.
- $\mathbb{E}[\text{Var}[Y|X]]$: average residual variance — the "unexplained" variance.
- This decomposition underpins $R^2$: the fraction of the total variance of $Y$ explained by $X$.
- You should be able to prove and apply both the LIE and LTV.

# Example: CEF from a Joint Density

Joint density:

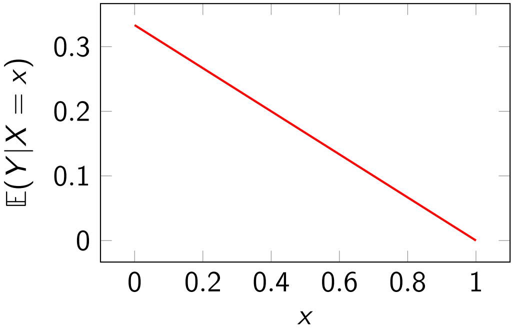
$$f(x,y) = 6(1-x-y)$$

for $x \geq 0,\ y \geq 0,\ x+y \leq 1.$

Marginal: $f_X(x) = 3(1-x)^2$

CEF:

$$\mathbb{E}(Y|X=x) = \frac{1-x}{3}$$




As $X$ increases, $\mathbb{E}[Y|X]$ decreases linearly—this CEF *is* linear.

# Linear Systems and Matrix Representation

A system of linear equations

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots \\
&\vdots
\end{aligned}
$$

can be written compactly as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$$

where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{X}$ is $n \times k$, and $\boldsymbol{\beta}$ is $k \times 1$.

# Matrix Multiplication

If $\boldsymbol{A}$ is $k \times r$ and $\boldsymbol{B}$ is $r \times s$, they are **conformable** and

$$(\boldsymbol{AB})_{ij} = \sum_{\ell=1}^{r} a_{i\ell}\, b_{\ell j}$$

The result is $k \times s$.

Key rules:

- Generally $\boldsymbol{AB} \neq \boldsymbol{BA}$.
- $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC} \quad$ (distributive).
- $(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC}) \quad$ (associative).
- $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$, where $'$ is the transpose.

# Inner Product and Similarity

- The inner product of two $k \times 1$ vectors:

$$\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}'\boldsymbol{b} = \sum_{j=1}^{k} a_j b_j$$

- Compare to covariance for demeaned variables:

$$\text{Cov}(\boldsymbol{x}, \ \boldsymbol{y}) = \frac{1}{n-1} \sum_{i=1}^{n} x_i y_i$$

- Two vectors are **orthogonal** if $\boldsymbol{a}'\boldsymbol{b} = 0$.
- $||\boldsymbol{a}|| = \sqrt{\boldsymbol{a}'\boldsymbol{a}}$ is the Euclidean norm (length) of $\boldsymbol{a}$.

# The Design Matrix $\boldsymbol{X}$

In regression, $\boldsymbol{X}$ is the $n \times k$ **design matrix**: rows are observations, columns are variables.

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,k-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,k-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} = \begin{bmatrix} \boldsymbol{i} & \boldsymbol{c}_1 & \boldsymbol{c}_2 & \cdots & \boldsymbol{c}_{k-1} \end{bmatrix}$$

- The first column $\boldsymbol{i} = (1, 1, \ldots, 1)'$ is the intercept.
- Each row $\boldsymbol{x}_i'$ is observation $i$'s vector of regressors.
- Each column $\boldsymbol{c}_j$ is the $n \times 1$ vector of all observations on variable $j$.
- $\boldsymbol{X}$ has $n$ rows (observations) and $k$ columns (parameters).

# Matrix Inverse

- If a $k \times k$ matrix $\boldsymbol{A}$ is *nonsingular* (full rank), there exists a unique $\boldsymbol{A}^{-1}$ such that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k$.
- Key formulas:
  - $(\boldsymbol{A}^{-1})' = (\boldsymbol{A}')^{-1}$
  - $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$
- The *rank* of a matrix is the number of linearly independent columns.
- A matrix is singular (non-invertible) when its columns are linearly dependent—this corresponds to **perfect multicollinearity** in regression.

# Quadratic Forms: An Example

- We often want to characterize a 2nd degree polynomial like

$$3x_1^2 + 4x_2^2 + 9x_3^2 - 5x_1x_3.$$

- We can write it as a quadratic form:

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}.$$

- The diagonal elements of $\boldsymbol{A}$ are the coefficients on $x_i^2$.
- Cross terms are split across symmetric entries:

$$-5x_1x_3 \Rightarrow a_{13} = a_{31} = -\frac{5}{2}.$$

- Thus,

$$\boldsymbol{A} = \begin{bmatrix} 3 & 0 & -\frac{5}{2} \\ 0 & 4 & 0 \\ -\frac{5}{2} & 0 & 9 \end{bmatrix}, \qquad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

# Positive Definiteness

- A **quadratic form** is a scalar $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is a symmetric matrix:

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_i a_{ii}x_i^2 + 2\sum_{i<j} a_{ij}x_i x_j$$

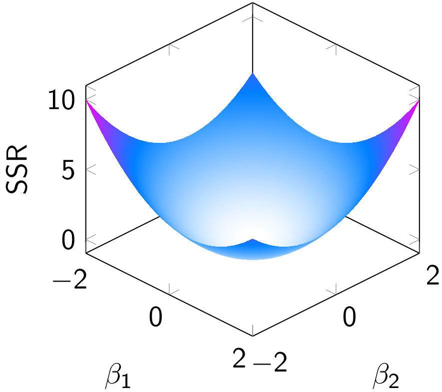
- A symmetric matrix $\boldsymbol{A}$ is **positive definite** if $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} > 0$ for all $\boldsymbol{c} \neq \mathbf{0}$.
- A symmetric matrix $\boldsymbol{A}$ is **positive semi-definite** if $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} \geq 0$ for all $\boldsymbol{c} \neq \mathbf{0}$.

# Why Positive Definiteness Matters

- Variance-covariance matrices are positive semi-definite by construction.
- If $\boldsymbol{X}$ has full column rank, $\boldsymbol{X}'\boldsymbol{X}$ is positive definite, guaranteeing a unique OLS solution.
- PD allows us to compare matrices (which estimator has "smaller" variance):

$$\boldsymbol{A} - \boldsymbol{B} \text{ is PD} \implies \boldsymbol{A} \text{ is "larger" than } \boldsymbol{B}$$

- A PD quadratic form is strictly convex with a unique global minimum—the OLS "bowl."



# Reference: Properties of Positive Definite Matrices

- $\boldsymbol{A}$ is PD $\iff$ it is symmetric with all eigenvalues positive.
- If $\boldsymbol{A}$ is PD, it is nonsingular and $\boldsymbol{A}^{-1}$ is also PD.
- If $\boldsymbol{A}$ is PD, $\text{tr}(\boldsymbol{A}) > 0$.
- If $\boldsymbol{A}$ and $\boldsymbol{B}$ are PD, so is $\boldsymbol{A} + \boldsymbol{B}$.
- If $\boldsymbol{A}$ is PD and $c > 0$, then $c\boldsymbol{A}$ is PD.
- If $\boldsymbol{A}$ is $n \times k$ with full column rank, then $\boldsymbol{A}'\boldsymbol{A}$ is PD.

# The Gram Matrix $X'X$

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} \sum x_{1i}^2 & \sum x_{1i}x_{2i} & \cdots \\ \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

- $\boldsymbol{X}'\boldsymbol{X}$ is symmetric and positive semi-definite.
- If we normalize columns so $||\boldsymbol{x}_j|| = 1$, then $(\boldsymbol{X}'\boldsymbol{X})_{ij} = \cos\theta_{ij}$: a measure of similarity.
- $\boldsymbol{X}'\boldsymbol{X}$ encodes all the second-moment information about the regressors.

# Linear Combinations

Given vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$,
a **linear combination** is any vector of the form

$$c_1 \boldsymbol{x}_1 + \cdots + c_k \boldsymbol{x}_k.$$

**Example in $\mathbb{R}^2$:**

- One vector: all multiples lie on a line.
- Two non-collinear vectors: combinations fill the plane.

Linear combinations describe what vectors you can “build.”

# Span

The **span** of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ is

$$\text{span}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} = \{c_1\boldsymbol{x}_1 + \cdots + c_k\boldsymbol{x}_k\}.$$

It is the set of *all* vectors you can build from them.

**Geometric intuition:**

- One independent vector $\Rightarrow$ a line.
- Two independent vectors $\Rightarrow$ a plane.
- Three independent vectors in $\mathbb{R}^3$ $\Rightarrow$ all of $\mathbb{R}^3$.

# Linear Independence and Basis

Vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are **linearly independent** if

$$c_1 \boldsymbol{x}_1 + \cdots + c_k \boldsymbol{x}_k = \mathbf{0}$$

implies

$$c_1 = \cdots = c_k = 0.$$

Intuition:

- No vector can be written as a combination of the others.
- No redundancy.

If independent vectors span a space $\mathcal{V}$, they form a **basis**.

# Dimension and the Column Space

**Dimension**

The dimension of a space is the number of vectors in any basis.

**In Regression**

- The columns of $\boldsymbol{X}$ are your regressors.
- Their span is the **column space**.
- OLS projects $\boldsymbol{y}$ onto this space.
- Linear dependence $\Rightarrow$ no unique solution.

Full column rank = regressors are linearly independent.

# Projection: The Geometric Heart of OLS

- **Projection Theorem:** Let $\mathcal{W}$ be a subspace. There exists a unique $\hat{\boldsymbol{y}} \in \mathcal{W}$ closest to $\boldsymbol{y}$:

$$\hat{\boldsymbol{y}} = \text{proj}_{\mathcal{W}}(\boldsymbol{y})$$

- **Projection onto a line:** If $\mathcal{W} = \{c\,\boldsymbol{x} : c \in \mathbb{R}\}$, then

$$\hat{\boldsymbol{y}} = \frac{\boldsymbol{x}'\boldsymbol{y}}{\boldsymbol{x}'\boldsymbol{x}}\,\boldsymbol{x}$$

- **General case:** If $\mathcal{W}$ is the column space of $\boldsymbol{X}$, then

$$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \equiv \boldsymbol{P}\boldsymbol{y}$$

where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the **hat matrix**.

# Orthogonality of Residuals

The error $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to every column of $\boldsymbol{X}$:

$$\begin{aligned} \boldsymbol{X}'(\boldsymbol{y} - \hat{\boldsymbol{y}}) &= \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}) \\ &= \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{0} \end{aligned}$$

- This is the matrix form of the OLS first-order conditions.
- Geometrically: the residual vector is perpendicular to the column space of $\boldsymbol{X}$.
- This is why $\sum_i \hat{e}_i = 0$ when there is an intercept (residuals are orthogonal to the column of ones).

# Summarizing Matrices

There are several useful summaries of a matrix $\boldsymbol{A}$:

- **Trace:**

$$\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} A_{ii}$$

- **Determinant:** $\det(\boldsymbol{A})$ measures (signed) volume scaling; $\det(\boldsymbol{A}) = 0$ iff $\boldsymbol{A}$ is singular.
- **Rank:** $\mathrm{rank}(\boldsymbol{A})$ is the dimension of the column space.
- **Eigenvalues:** $\lambda_1, \ldots, \lambda_n$ summarize stretching along special directions.

We will use these repeatedly to diagnose invertibility and curvature.

# Properties of the Trace

The trace has several properties that we will use repeatedly:

1. $\text{tr}(\boldsymbol{A}+\boldsymbol{B}) = \text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B})$ (linearity)
2. $\text{tr}(c\boldsymbol{A}) = c\ \text{tr}(\boldsymbol{A})$ (linearity)
3. $\text{tr}(\boldsymbol{A}') = \text{tr}(\boldsymbol{A})$ (transpose invariance)
4. $\text{tr}(\boldsymbol{AB}) = \text{tr}(\boldsymbol{BA})$ (cyclic property)

**Proof of (4):** $\text{tr}(\boldsymbol{AB}) = \sum_i (\boldsymbol{AB})_{ii} = \sum_i \sum_j a_{ij} b_{ji} = \sum_j \sum_i b_{ji} a_{ij} = \sum_j (\boldsymbol{BA})_{jj} = \text{tr}(\boldsymbol{BA})$

**Where this matters:** When we prove that $s^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-k}$ is unbiased for $\sigma^2$, the key step uses

$$\mathbb{E}[\boldsymbol{e}'\boldsymbol{e}|\boldsymbol{X}] = \mathbb{E}[\text{tr}(\boldsymbol{M}\boldsymbol{e}\boldsymbol{e}')|\boldsymbol{X}] = \text{tr}(\boldsymbol{M}\,\mathbb{E}[\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}]) = \sigma^2\,\text{tr}(\boldsymbol{M}) = \sigma^2(n-k)$$

# Eigenvalues and Eigenvectors

- For a square matrix $\boldsymbol{A}$, if

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}$$

  for some nonzero $\boldsymbol{u}$, then $\boldsymbol{u}$ is an **eigenvector** and $\lambda$ is the corresponding **eigenvalue**.
- Interpretation: along direction $\boldsymbol{u}$, the matrix acts like multiplication by $\lambda$.
- If $\boldsymbol{A}$ is symmetric, its eigenvalues are real and it has an orthonormal eigenbasis.
- Two useful identities:

$$\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i \qquad \text{and} \qquad \det(\boldsymbol{A}) = \prod_{i=1}^{n} \lambda_i.$$

- Eigenvalues diagnose key properties (symmetric $\boldsymbol{A}$):
  - $\lambda_i > 0$ for all $i \iff \boldsymbol{A}$ is positive definite.
  - Some $\lambda_i = 0 \iff \boldsymbol{A}$ is singular.

# Idempotent Matrices

- A matrix is **idempotent** if
$$\boldsymbol{A}^2 = \boldsymbol{A}.$$

## Theorem

*If $\boldsymbol{A}$ is idempotent, then all eigenvalues of $\boldsymbol{A}$ are $0$ or $1$.*

- **Proof (one line):** If $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$, then
$$\boldsymbol{A}^2\boldsymbol{x} = \lambda^2\boldsymbol{x} \quad \text{but also} \quad \boldsymbol{A}^2\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x},$$
so $\lambda^2 = \lambda$, hence $\lambda \in \{0, 1\}$.
- The hat matrix
$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$
is symmetric and idempotent.
- $\text{rank}(\boldsymbol{P})$ equals the number of eigenvalues equal to 1 (so $\text{rank}(\boldsymbol{P}) = k$).
- Therefore $\text{tr}(\boldsymbol{P}) = k$.

# Orthogonal Complements and Dimension

## Definition

*The* **orthogonal complement** *of a subspace* $\mathcal{W} \subset \mathbb{R}^n$ *is*

$$\mathcal{W}^{\perp} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x}'\boldsymbol{w} = 0 \textit{ for all } \boldsymbol{w} \in \mathcal{W}\}.$$

## Theorem (Fundamental Theorem of Linear Algebra (dimension version))

*If* $\mathcal{W}$ *is a subspace of* $\mathbb{R}^n$, *then*

$$\dim(\mathcal{W}) + \dim(\mathcal{W}^{\perp}) = n.$$

**Proof setup:** Let $\dim(\mathcal{W}) = k$. Choose an orthonormal basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ for $\mathcal{W}$. Extend it to an orthonormal basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ for $\mathbb{R}^n$.

# Proof: $\dim(\mathcal{W}) + \dim(\mathcal{W}^\perp) = n$

**Claim:** $\boldsymbol{u}_{k+1}, \ldots, \boldsymbol{u}_n$ form a basis for $\mathcal{W}^\perp$, so $\dim(\mathcal{W}^\perp) = n - k$.

**Step 1 (orthogonality):** For $j > k$ and $i \leq k$, orthonormality gives $\boldsymbol{u}_j'\boldsymbol{u}_i = 0$, hence $\boldsymbol{u}_j \in \mathcal{W}^\perp$.

**Step 2 (spanning):** Take any $\boldsymbol{v} \in \mathcal{W}^\perp$ and expand in the full basis:

$$\boldsymbol{v} = \sum_{i=1}^{n} c_i \boldsymbol{u}_i.$$

For any $j \leq k$,

$$0 = \boldsymbol{v}'\boldsymbol{u}_j = c_j,$$

so $\boldsymbol{v} = \sum_{i=k+1}^{n} c_i \boldsymbol{u}_i$, which lies in $\operatorname{span}\{\boldsymbol{u}_{k+1}, \ldots, \boldsymbol{u}_n\}$.

**Step 3 (independence):** $\boldsymbol{u}_{k+1}, \ldots, \boldsymbol{u}_n$ are orthonormal, hence linearly independent.

Therefore $\dim(\mathcal{W}^\perp) = n - k$, so $\dim(\mathcal{W}) + \dim(\mathcal{W}^\perp) = k + (n - k) = n$. $\square$

# Degrees of Freedom

Let $\mathcal{W} = \text{col}(\boldsymbol{X})$ be the column space of $\boldsymbol{X}$.

- $\dim(\mathcal{W}) = k$ (number of linearly independent regressors).
- The fitted values satisfy $\hat{\boldsymbol{y}} \in \mathcal{W}$.
- The residuals satisfy $\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}} \in \mathcal{W}^{\perp}$.
- By $\dim(\mathcal{W}) + \dim(\mathcal{W}^{\perp}) = n$,

$$\dim(\mathcal{W}^{\perp}) = n - k.$$

- This is why we divide by $n - k$ when estimating $\sigma^2$: residual variation lives in an $(n-k)$-dimensional space.

# What Is the Derivative of a Vector Function?

Let $f(a_1, a_2, a_3, \ldots a_k) = f(\boldsymbol{a})$ be a scalar function of a vector $\boldsymbol{a} \in \mathbb{R}^k$.

The **gradient** is:

$$\nabla_{\boldsymbol{a}} f(\boldsymbol{a}) = \begin{bmatrix} \dfrac{\partial f}{\partial a_1} \\ \vdots \\ \dfrac{\partial f}{\partial a_k} \end{bmatrix}.$$

**Key idea:**

- Derivative of a scalar w.r.t. a vector is a vector.
- First-order condition for a minimum:

$$\nabla_{\boldsymbol{a}} f(\boldsymbol{a}) = \mathbf{0}.$$

# Linear Forms

Let $f(\boldsymbol{a}) = \boldsymbol{z}'\boldsymbol{a}$, where $\boldsymbol{z}$ is fixed.

Write it out:

$$f(\boldsymbol{a}) = \sum_{j=1}^{k} z_j a_j.$$

Taking derivatives componentwise:

$$\nabla_{\boldsymbol{a}}(\boldsymbol{z}'\boldsymbol{a}) = \boldsymbol{z}.$$

More generally, if $f(\boldsymbol{a}) = \boldsymbol{Z}\boldsymbol{a}$,

$$\frac{d\,\boldsymbol{Z}\boldsymbol{a}}{d\,\boldsymbol{a}} = \boldsymbol{Z}.$$

# Derivatives of Quadratic Forms

Let

$$f(\boldsymbol{a}) = \boldsymbol{a}'\boldsymbol{Z}\boldsymbol{a}.$$

Write it out:

$$f(\boldsymbol{a}) = \sum_{i,j} a_i Z_{ij} a_j.$$

Taking derivatives:

$$\nabla_{\boldsymbol{a}}(\boldsymbol{a}'\boldsymbol{Z}\boldsymbol{a}) = (\boldsymbol{Z} + \boldsymbol{Z}')\boldsymbol{a}.$$

If $\boldsymbol{Z}$ is symmetric, so $\boldsymbol{Z} = \boldsymbol{Z}'$ :

$$\nabla_{\boldsymbol{a}}(\boldsymbol{a}'\boldsymbol{Z}\boldsymbol{a}) = 2\boldsymbol{Z}\boldsymbol{a}.$$

## Application: Deriving OLS

Least squares solves:

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta).$$

Expand:

$$= \boldsymbol{y}'\boldsymbol{y} - 2\beta'\boldsymbol{X}'\boldsymbol{y} + \beta'\boldsymbol{X}'\boldsymbol{X}\beta.$$

Take gradient and set equal to zero:

$$-2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\beta = \mathbf{0}.$$

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

# Second-Order Condition: Verifying a Minimum

The FOC gave us a critical point. Is it a minimum?

The **Hessian** (matrix of second derivatives) of $(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$ is:

$$\frac{\partial^2}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'} \left[\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\beta\right] = 2\boldsymbol{X}'\boldsymbol{X}$$

- If $\boldsymbol{X}$ has full column rank, then $\boldsymbol{X}'\boldsymbol{X}$ is positive definite (from our earlier result).
- A positive definite Hessian means the objective is strictly convex — the critical point is a unique global minimum.
- This connects two sections: positive definiteness guarantees both that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists *and* that the solution is a minimum.

# Summary and Roadmap

- The **CEF** $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is the target of regression—the best MSE predictor.
- **OLS** is the linear approximation: a projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$.
- The OLS residual is orthogonal to $\boldsymbol{X}$ (first-order conditions).
- Positive definiteness of $\boldsymbol{X}'\boldsymbol{X}$ guarantees existence and uniqueness.
- Degrees of freedom $(n - k)$ come from the rank-nullity theorem.

- **Next:** the CEF vs Best Linear Predictor