Math Revision Session

3: Matrix Algebra (3)

Jukina HATAKEYAMA

The University of Osaka, Department of Economics

April 6, 2026

Solving a System of Linear Equations with Matrices

Step 1: Write the system in matrix form

Consider

$$\begin{cases} 2x + y = 5, \\ 3x + 4y = 6. \end{cases}$$

This system can be written as

$$A\mathbf{x} = \mathbf{b},$$

where

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 5 \\ 6 \end{pmatrix}.$$

What Does the Existence of A^{-1} Mean?

If A^{-1} exists, then

$$A^{-1}A = AA^{-1} = I.$$

This means that the linear transformation by A can be **undone**.

- A transforms x into Ax.
- A^{-1} transforms Ax back into x.
- So A^{-1} is not just "division by a matrix".
- It is the matrix that **reverses** the transformation represented by A.

If A^{-1} does not exist, then some information is lost, and the original vector cannot always be recovered uniquely.

Solving the System using the Inverse Matrix

Step 2: Use A^{-1}

If A is invertible, then

$$\mathbf{x} = A^{-1}\mathbf{b}.$$

For

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix},$$

we have

$$\det(A) = 2 \cdot 4 - 1 \cdot 3 = 5 \neq 0.$$

Hence,

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} 4 & -1 \\ -3 & 2 \end{pmatrix} = \frac{1}{5} \begin{pmatrix} 4 & -1 \\ -3 & 2 \end{pmatrix}.$$

Computing the Solution

Step 3: Compute $A^{-1}\mathbf{b}$

$$\mathbf{x} = \frac{1}{5}\begin{pmatrix} 4 & -1 \\ -3 & 2 \end{pmatrix}\begin{pmatrix} 5 \\ 6 \end{pmatrix}$$

$$= \frac{1}{5}\begin{pmatrix} 4 \cdot 5 + (-1) \cdot 6 \\ (-3) \cdot 5 + 2 \cdot 6 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 14 \\ -3 \end{pmatrix}.$$

Therefore,

$$\mathbf{x} = \begin{pmatrix} 14/5 \\ -3/5 \end{pmatrix} = \begin{pmatrix} 2.8 \\ -0.6 \end{pmatrix}.$$

Cramer's Rule

For a square system

$$A\mathbf{x} = \mathbf{b},$$

if $\det(A) \neq 0$, then each variable can be written as

$$x_i = \frac{\det(A_i)}{\det(A)},$$

where A_i is the matrix obtained by replacing the i-th column of A with $\mathbf{b}$.

Remark:

- Cramer's Rule is useful for small systems.
- For large systems, matrix methods such as Gaussian elimination are more practical.

Example of Cramer's Rule

Consider

$$\begin{cases} x + y + z = 6, \\ 2x + 3y + z = 14, \\ 3x + y + 2z = 10. \end{cases}$$

Then

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 6 \\ 14 \\ 10 \end{pmatrix}.$$

We first compute $\det(A)$.

Computing the Determinant

$$\det(A) = 1(3 \cdot 2 - 1 \cdot 1) - 1(2 \cdot 2 - 1 \cdot 3) + 1(2 \cdot 1 - 3 \cdot 3)$$

$$= 1(6 - 1) - 1(4 - 3) + 1(2 - 9)$$

$$= 5 - 1 - 7 = -3.$$

Since $\det(A) \neq 0$, the system has a unique solution.

Replacing the Columns

The matrices for Cramer's Rule are

$$A_x = \begin{pmatrix} 6 & 1 & 1 \\ 14 & 3 & 1 \\ 10 & 1 & 2 \end{pmatrix}, \qquad A_y = \begin{pmatrix} 1 & 6 & 1 \\ 2 & 14 & 1 \\ 3 & 10 & 2 \end{pmatrix}, \qquad A_z = \begin{pmatrix} 1 & 1 & 6 \\ 2 & 3 & 14 \\ 3 & 1 & 10 \end{pmatrix}.$$

Their determinants are

$$\det(A_x) = -4, \qquad \det(A_y) = -10, \qquad \det(A_z) = -4.$$

Final Solution by Cramer's Rule

Using

$$x_i = \frac{\det(A_i)}{\det(A)},$$

we obtain

$$x = \frac{-4}{-3} = \frac{4}{3}, \qquad y = \frac{-10}{-3} = \frac{10}{3}, \qquad z = \frac{-4}{-3} = \frac{4}{3}.$$

Therefore,

$$\boxed{x = \frac{4}{3}, \qquad y = \frac{10}{3}, \qquad z = \frac{4}{3}}$$

Matrix Differentiation

In econometrics and statistics, matrix differentiation is important because many objective functions are written in matrix form.

Examples include:

- least squares,
- maximum likelihood,
- quadratic forms,
- optimisation problems with vectors and matrices.

Today, we focus on derivatives with respect to a vector, since this is what we need for OLS.

Gradient

Let $f(x)$ be a scalar-valued function of

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

The gradient of $f(x)$ with respect to x is

$$\nabla_x f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}.$$

It gives the direction of the steepest increase of the function.

Some Useful Derivatives

Let $a \in \mathbb{R}^n$, $x \in \mathbb{R}^n$, and $A \in \mathbb{R}^{n \times n}$. Then:

$$\frac{\partial}{\partial x}(a^\top x) = a,$$

$$\frac{\partial}{\partial x}(x^\top a) = a,$$

$$\frac{\partial}{\partial x}(x^\top A x) = (A + A^\top)x.$$

If A is symmetric, this becomes

$$\frac{\partial}{\partial x}(x^\top A x) = 2Ax.$$

Also,

$$\frac{\partial}{\partial x}(x^\top x) = 2x.$$

Derivative with Respect to a Matrix

Let $A = (a_{ij}) \in \mathbb{R}^{m \times n}$, and let $f(A)$ be a scalar-valued function of A. Then the derivative of $f(A)$ with respect to A is the matrix

$$\frac{\partial f(A)}{\partial A} = \begin{pmatrix} \frac{\partial f(A)}{\partial a_{11}} & \cdots & \frac{\partial f(A)}{\partial a_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial a_{m1}} & \cdots & \frac{\partial f(A)}{\partial a_{mn}} \end{pmatrix}.$$

For example, if

$$f(A) = \mathrm{tr}(A^\top A) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2,$$

then

$$\frac{\partial f(A)}{\partial A} = 2A.$$

Chain Rule

The chain rule extends naturally to multivariate functions.

Suppose that $f(g(x))$ is a composite function. Then we differentiate the outer function and multiply by the derivative of the inner function.

In matrix notation, this idea is often written using gradients or Jacobian matrices.

In this lecture, we mainly use the chain rule when differentiating quadratic functions such as

$$(y - X\beta)^\top (y - X\beta).$$

A Simple Example of the Chain Rule

Let

$$u(x) = Ax - b,$$

and define

$$f(x) = u(x)^\top u(x).$$

Then

$$f(x) = (Ax - b)^\top (Ax - b).$$

Using the chain rule and the derivative of a quadratic form, we obtain

$$\nabla_x f(x) = 2A^\top (Ax - b).$$

This type of calculation appears frequently in least squares problems.

Ordinary Least Squares

In ordinary least squares (OLS), we estimate β by minimising the sum of squared residuals.

For observation i,

$$y_i = x_i^\top \beta + \varepsilon_i,$$

where

- y_i is the dependent variable,
- $x_i \in \mathbb{R}^k$ is the regressor vector,
- $\beta \in \mathbb{R}^k$ is the coefficient vector,
- ε_i is the error term.

The OLS estimator solves

$$\min_\beta \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2.$$

OLS in Matrix Form

Stacking all observations, we write

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

where

$$\mathbf{y}, \varepsilon \in \mathbb{R}^n, \qquad \mathbf{X} \in \mathbb{R}^{n \times k}, \qquad \beta \in \mathbb{R}^k.$$

Then the objective function becomes

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 = \min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta).$$

This is a quadratic minimisation problem in β.

Expanding the Objective Function

Define

$$Q(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta).$$

Expanding,

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

So we only need derivatives of:

- a linear term $\beta^\top a$,
- a quadratic term $\beta^\top A\beta$.

Differentiating the OLS Objective

Using

$$\frac{\partial}{\partial \beta}(\beta^\top a) = a$$

and

$$\frac{\partial}{\partial \beta}(\beta^\top A \beta) = (A + A^\top)\beta,$$

we obtain

$$\nabla_\beta Q(\beta) = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\beta,$$

because $\mathbf{X}^\top \mathbf{X}$ is symmetric.

First-Order Condition

Setting the gradient equal to zero gives

$$\nabla_\beta Q(\beta) = 0$$

and hence

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

Therefore,

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

This is called the **normal equation**.

OLS Estimator

If $\mathbf{X}^\top\mathbf{X}$ is invertible, then the solution is

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

Important:

- $\mathbf{X}^\top\mathbf{X}$ is invertible when the columns of $\mathbf{X}$ are linearly independent.
- So rank is important here: if $\mathrm{rank}(\mathbf{X}) < k$, then $\mathbf{X}^\top\mathbf{X}$ is singular.

This is one of the most important applications of matrix algebra in econometrics.

Gradient, Jacobian, and Hessian

These three concepts are all based on derivatives, but they are used for different types of functions.

- **Gradient:** derivative of a scalar-valued function with respect to a vector.
- **Jacobian:** matrix of first derivatives of a vector-valued function.
- **Hessian:** matrix of second derivatives of a scalar-valued function.

Gradient

If $f(x)$ is a scalar-valued function of $x \in \mathbb{R}^n$, then its gradient is

$$\nabla_x f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}.$$

Example: If

$$f(x) = x^\top A x$$

and A is symmetric, then

$$\nabla_x f(x) = 2Ax.$$

Jacobian

If

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{pmatrix}, \qquad x \in \mathbb{R}^n,$$

then the Jacobian matrix is

$$J_f(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

It collects all first-order partial derivatives of a vector-valued function.

Hessian

If $f(x)$ is a scalar-valued function of $x \in \mathbb{R}^n$, then the Hessian matrix is

$$H_f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}.$$

It describes the curvature of the function and is useful in optimisation.

Assumptions for the Variance of $\hat{\beta}$

To derive the variance of the OLS estimator, we usually assume:

- $y = X\beta + \varepsilon$,
- $E[\varepsilon \mid X] = 0$,
- $\mathrm{Var}(\varepsilon \mid X) = \sigma^2 I_n$.

The last assumption means:

- the error variance is constant;
- the errors are uncorrelated across observations.

Deriving the Variance of the OLS Estimator

Recall that

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

Since $y = X\beta + \varepsilon$, we have

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

Therefore,

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

Variance of the OLS Estimator

Using

$$\hat{\beta} - \beta = (X^\top X)^{-1} X^\top \varepsilon,$$

we obtain

$$\mathrm{Var}(\hat{\beta} \mid X) = \mathrm{Var}\Big((X^\top X)^{-1} X^\top \varepsilon \,\Big|\, X\Big).$$

Since $(X^\top X)^{-1} X^\top$ is non-random conditional on X,

$$\mathrm{Var}(\hat{\beta} \mid X) = (X^\top X)^{-1} X^\top \,\mathrm{Var}(\varepsilon \mid X) X (X^\top X)^{-1}.$$

If $\mathrm{Var}(\varepsilon \mid X) = \sigma^2 I_n$, then

$$\mathrm{Var}(\hat{\beta} \mid X) = \sigma^2 (X^\top X)^{-1}.$$

What Does This Mean?

The covariance matrix

$$\mathrm{Var}(\hat{\beta} \mid X) = \sigma^2 (X^\top X)^{-1}$$

tells us how precisely the OLS estimator is estimated.

- The diagonal elements are the variances of each component of $\hat{\beta}$.
- The off-diagonal elements are the covariances between components.
- Smaller variances mean more precise estimates.

So the covariance matrix is essential for:

- standard errors,
- confidence intervals,
- hypothesis testing.

Estimating the Variance in Practice

In practice, σ^2 is unknown, so we estimate it by

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^\top \hat{\varepsilon}}{n-k},$$

where

$$\hat{\varepsilon} = y - X\hat{\beta}.$$

Then the estimated covariance matrix of $\hat{\beta}$ is

$$\widehat{\mathrm{Var}}(\hat{\beta} \mid X) = \hat{\sigma}^2 (X^\top X)^{-1}.$$

The standard error of $\hat{\beta}_j$ is the square root of the j-th diagonal element of this matrix.

Gradient and Hessian of the OLS Objective

Recall the OLS objective:

$$Q(\beta) = \frac{1}{2}(y - X\beta)^\top (y - X\beta).$$

Define the gradient by

$$g(\beta) := \nabla_\beta Q(\beta).$$

Then

$$g(\beta) = -X^\top y + X^\top X\beta.$$

Define the Hessian by

$$H(\beta) := \nabla^2_\beta Q(\beta).$$

Then

$$H(\beta) = X^\top X.$$

Connection with the OLS Estimator

The first-order condition is

$$g(\beta) = 0.$$

Using

$$g(\beta) = -X^\top y + X^\top X\beta,$$

we obtain

$$X^\top X\beta = X^\top y.$$

Hence,

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

So the gradient $g(\beta)$ determines the first-order condition.

Connection with the Covariance Matrix

The Hessian is

$$H(\beta) = X^\top X.$$

Therefore,

$$\mathrm{Var}(\hat{\beta} \mid X) = \sigma^2 (X^\top X)^{-1} = \sigma^2 H(\beta)^{-1}.$$

So:

- $g(\beta)$ is used to derive the estimator;
- $H(\beta)$ describes the curvature of the objective function;
- $H(\beta)$ also determines the precision of $\hat{\beta}$.