# Linear Algebra for Business Analytics

Marcel Scharth
The University of Sydney Business School

This version: 11/8/2017

This reference material for Business Analytics students covers basic concepts from linear algebra that are helpful in applications. You can use this guide either to learn the essentials, or as a reference that you can always go back to as you come across these concepts. This material draws on Klein (2013) and Boyd and Vandenberghe (2016), with the latter being particularly well suited as a complete resource for business analytics applications.

Here, we follow a practical approach and do not cover abstract linear algebra, which includes concepts that would be part of a traditional linear algebra course such as vector spaces. While powerful, this level of abstraction is not a requirement for our purposes. Instead, you will find that investing in the material below considerably simplifies things from the perspective of understanding practical methods.

## Contents

The † symbol indicates subsections that are less important and can be initially skipped. You can always go back to them when you come across these concepts.

# 1. Vectors

## 1.1. What is a vector?

A **vector** is an ordered finite list of numbers. We typically write vectors as vertical arrays, surrounded by square or curved brackets, as in

$$\begin{bmatrix} -1 \\ 0 \\ 2.5 \\ -7.2 \end{bmatrix} \text{ or } \begin{pmatrix} -1 \\ 0 \\ 2.5 \\ -7.2 \end{pmatrix}$$

$$\mathbf{a} = \begin{bmatrix} 5 \\ -2 \\ -3 \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

We also write vectors as numbers separated by commas.

$$\mathbf{a} = (5, -2, -3), \quad \mathbf{a} = (1, 1).$$

The **elements** or **entries** of a vector are the values in the array.

The **size** (or **dimension**) of a vector is the number of elements it contains.

A vector of size $n$ is called an $n-$vector.

$$\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_n \end{bmatrix}$$

$$\mathbf{a} = (a_1, \cdots, a_i, \cdots, a_n)$$

A vector with $n$ entries, each belonging to $\mathbb{R}$ (the set of real numbers), is called a $n$-vector over $\mathbb{R}$. We denote the set of $n$-vectors over $\mathbb{R}$ as $\mathbb{R}^n$.

$$\boldsymbol{a} = (a_1, a_2, a_3) \in (\mathbb{R}, \mathbb{R}, \mathbb{R}) \equiv \mathbb{R}^3$$

We can also define a vector as a function from a finite set $D$ to $\mathbb{R}$. For example, a function from $D = \{0, 1, 2, \ldots, d-1\}$ to $\mathbb{R}$. The vector

$$\mathbf{a} = (6, -4, -3.7)$$

is the function

$$\begin{aligned} 0 &\longmapsto 6 \\ 1 &\longmapsto -4 \\ 2 &\longmapsto -3.7, \end{aligned}$$

where $\longmapsto$ reads "maps to".

This last definition is useful as it matches how we work with vectors in Python. For example, if we store the above vector as a Python list called `a`, the command `a[0]` returns 6, the first element of the vector. More formally, we say that the above definition lends itself to representation in a data structure (a format for organising and storing data). Python objects such as lists, dictionaries, and `NumPy` arrays are data structures that can represent vectors.

### 1.2. **Special vectors.**

**Zero vector**. A zero vector has all elements equal to zero $\mathbf{0} = (0, 0, ..., 0)$. $\mathbf{0}_n$ indicates a zero vector with dimension $n$.

**Unit vector**. A unit vector has all elements equal zeros, except one element which is equal to one. $\mathbf{e}_i = [0, ..., 0, 1, 0, ..., 0]^T$ (all zeros except for 1 at $i$-th position).

**Ones vector**. A ones vector has elements equal to one, $\mathbf{1} = [1, 1, ..., 1]^T$. $\mathbf{1}_n$ indicates a ones vector with dimension $n$. We also use the notation $\boldsymbol{\iota}$ for this type of vector

**Sparsity**. A vector is said to be sparse if many of its elements are equal to zero.

## 1.3. **Vector operations.**

**Vector equality**. $\mathbf{a} = \mathbf{b} \iff a_i = b_i$ for all $i = 1, 2, ..., n$.

**Scalar-vector multiplication**. Let $\alpha$ denote a scalar. The vector $\alpha\,\mathbf{a}$ is the vector with elements $\{\alpha\, a_i\}$. For example, let $\mathbf{a} = (5, -2, -3)$, then

$$0.5\,\mathbf{a} = (0.5 \times 5, 0.5 \times -2, 0.5 \times -3) = (2.5, -1, -1.5)$$

**Addition**. let $\mathbf{a}$ and $\mathbf{b}$ be two vectors with the same size $n$. The sum $\mathbf{c} = \mathbf{a} + \mathbf{b}$ is the vector with elements $c_i = a_i + b_i$.

Let $\mathbf{a} = (5, -2, -3)$ and $\mathbf{b} = (-1, 2, 4)$. Then,

$$\mathbf{c} = \mathbf{a} + \mathbf{b} = (5, -2, -3) \; + \; (-1, 2, 4) = (5 - 1, -2 + 2, -3 + 4) = (4, 0, 1).$$

**Linear combination**. Let $\mathbf{a}$ and $\mathbf{b}$ are $n-$vectors and $\beta_1$ and $\beta_2$ are scalars, the $n-$vector

$$\beta_1\mathbf{a} + \beta_2\mathbf{b}$$

is called a linear combination of $\boldsymbol{a}$ and $\boldsymbol{b}$. The scalars $\beta_1$ and $\beta_2$ are the **coefficients** of the linear combination.

Let $\mathbf{a} = (5, -2, -3)$, $\mathbf{b} = (-1, 2, 4)$, $\beta_1 = 2$, and $\beta_2 = 3$.

$$2\mathbf{a} + 3\mathbf{b} = (2 \times 5, 2 \times -2, 2 \times -3) + (3 \times -1, 3 \times 2, 3 \times 4) = (7, 2, 18)$$

## 1.4. **Inner Product.**

We define the dot or **inner product** of two $n$-dimensional vectors $\mathbf{a}$ and $\mathbf{b}$ as

$$\boldsymbol{a}^T\boldsymbol{b} = a_1b_1 + a_2b_2 + \ldots + a_nb_n = \sum_{i=1}^{n} a_ib_i$$

Example: $\mathbf{a} = (2, -1, 3)$ and $\mathbf{b} = (5, -2, -3)$, then

$$\boldsymbol{a}^T\boldsymbol{b} = 2 \times 5 + (-1) \times (-2) + 3 \times (-3) = 3$$

Some authors use the notation $\langle \mathbf{a}, \mathbf{b} \rangle$ for inner products.

**Properties**. The following are useful properties of inner products that follow easily from the definition.

$$(\alpha \mathbf{a})^T \mathbf{b} = \alpha(\mathbf{a}^T \mathbf{b})$$

$$\mathbf{a}^T \mathbf{b} = \mathbf{b}^T \mathbf{a}$$

$$\mathbf{a}^T(\mathbf{b} + \mathbf{c}) = \mathbf{a}^T\mathbf{b} + \mathbf{a}^T\mathbf{c}$$

.

**Examples**.

Sum. $\boldsymbol{\iota}^T\boldsymbol{a} = a_1 + a_2 + \ldots + a_n$ is the sum the elements of $\boldsymbol{a}$.

Average. $(1/n)(\boldsymbol{\iota}^T\boldsymbol{a})$ is the average of the elements of $\boldsymbol{a}$.

Sum of squares. $\boldsymbol{a}^T\boldsymbol{a} = a_1^2 + \ldots + a_n^2$ is the sum of squares of the elements of $\boldsymbol{a}$.

$$n = 4, \quad \mathbf{x} = \begin{bmatrix} 3 \\ 4 \\ 2 \\ 7 \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

$$\Rightarrow \frac{1}{4}\boldsymbol{\iota}^T\mathbf{x} = \frac{1}{4}(1 \times 3 + 1 \times 4 + 1 \times 3 + 1 \times 7) = 4.$$

1.5. **Linear functions.**

The notation $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ means that $f$ is a function that maps an $n-$vector to a real number. If $\boldsymbol{x}$ is an $n-$vector, then $f(\boldsymbol{x})$ (a scalar) is the value of the function at $\boldsymbol{x}$. In this setting, we refer to $\boldsymbol{x}$ as the **argument** of the function.

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be $n-$vectors and $\alpha$ and $\beta$ be scalars. A **linear function** is a function that satisfies the property

$$f(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha f(\boldsymbol{x}) + \beta f(\boldsymbol{y})$$

We can always represent a linear function as an inner product. Let $\boldsymbol{a}$ be an $n-$vector. The we can write any linear function using the form

$$f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} = a_1x_1 + a_2x_2 + \ldots a_nx_n,$$

where $\boldsymbol{x}$ is an $n-$vector. Here, $\boldsymbol{a}$ is fixed, and the argument $\boldsymbol{x}$ can be any $n-$vector.

For example, in a linear regression model $f(\boldsymbol{x})$ is the regression function, $\boldsymbol{x}$ are the predictor values, and $\boldsymbol{a}$ are the model parameters.

An **affine function** is a linear function plus a constant, that is

$$f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} + b,$$

for a scalar $b$.

1.6. **Norm.**

The **Euclidean norm** or $\ell_2$-norm of a vector is

$$\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^T\mathbf{a}} = \left(\sum_{i=1}^{n} a_i^2\right)^{1/2}.$$

This is the distance from the origin to the point $\mathbf{a}$ or the **length** of the vector.

The **normalized vector** $\frac{\mathbf{a}}{\|\mathbf{a}\|_2}$ has unit norm.

**Example.**

Let $\mathbf{x}$ be a vector with sample average zero. Then $\|\mathbf{x}\|^2$ is the sum of squares of $\boldsymbol{x}$ and $s_x = \|\mathbf{x}\|^2/n$ is the sample variance.

**General definition.** A norm $\|\cdot\|$ is a function that satisfies the following properties:

(1) $\|\boldsymbol{a}\| \geq 0$ (non-negativity).
(2) $\|\boldsymbol{a}\| = 0$ only if $\boldsymbol{x} = \mathbf{0}$ (definiteness).
(3) $\|\alpha\boldsymbol{a}\| = |\alpha| \times \|\boldsymbol{a}\|$ (homogeneity).
(4) $\|\boldsymbol{a} + \boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$ (triangle inequality).

The $\ell_1$ norm of a vector is

$$\|\mathbf{a}\|_1 = |a_1| + |a_2| + \ldots + |a_n| = \sum_{i=1}^{n} |a_i|.$$

The **Chebyshev** or $\ell_\infty$ norm is given by

$$\|\mathbf{a}\|_\infty = \max\{|a_1|, |a_2|, \ldots, |a_n|\}.$$

The **Minkowski norm** of order $p$ is

$$\|\mathbf{a}\|_p = \left(\sum_{i=1}^{n} |a_i|^p\right)^{1/p},$$

for $p \geq 1$. This is a generalisation of the previous norms. We include this here because the `scikit-learn` package in Python sometimes refers to the Minkowski norm by default, even if $p = 2$.

### 1.7. **Distance.**

The **Euclidean distance** between two vectors $\mathbf{x}$ and $\mathbf{y}$ is the norm of the difference vector $\mathbf{x} - \mathbf{y}$:

$$\mathbf{dist}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2 = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y})} = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{1/2}$$

Every norm $\|\cdot\|$ induces a distance metric $\|\boldsymbol{x}-\boldsymbol{y}\|$.

### 1.8. **Orthogonal vectors.**

Two vectors are **orthogonal**, written $\mathbf{a} \perp \mathbf{b}$, if and only if their inner product is zero, $\mathbf{a}^T\mathbf{b} = 0$.

Example:

$$\mathbf{a} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \|\mathbf{a}\| = \|\mathbf{b}\| = \sqrt{2}, \ \mathbf{a}^T\mathbf{b} = 0.$$

$$\mathbf{c} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} .6 \\ .2 \end{bmatrix},$$

$$\|\mathbf{c}\| = \sqrt{10} \approx 3.16, \|\mathbf{d}\| = \sqrt{.4} \approx 0.63, \quad \mathbf{c}^T\mathbf{d} = 0, \text{ and } \mathbf{a}^T\mathbf{c} = -2.$$

## 2. Matrices

### 2.1. **What is a matrix?**

A matrix is a rectangular two-dimensional array of numbers such as

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & -2.3 & 0.1 \\ 1.3 & 4 & -0.1 & 7 \\ 4.1 & -1 & 0 & 1.7 \end{bmatrix}$$

The **size** (or **dimensions**) of a matrix are the number of rows and columns. The matrix above has 3 rows and 4 columns, so the size is $3 \times 4$ (it reads 3-by-4).

We represent an $(m \times n)$ matrix as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn}, \end{bmatrix}$$

with $\mathbf{A} \in \mathbb{R}^{m \times n}$.

We also represent a matrix as $\mathbf{A} = \{a_{ij}\}$. In a design matrix in regression analysis, the index $i = 1, 2, ..., m$ refers to the statistical units, and the index $j = 1, 2, ..., n$ to the variables or attributes.

## 2.2. **Row and column vectors.**

A column vector is a $m \times 1$ matrix. We do not distinguish between vectors and column vectors.

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

In the same way, a row vector is a $1 \times m$ matrix.

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \dots & a_m \end{bmatrix}$$

The **transpose** of a column vector is the corresponding row vector and vice-versa.

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}^T = \begin{bmatrix} a_1 & a_2 & \dots & a_m \end{bmatrix}$$

$$\begin{bmatrix} a_1 & a_2 & \dots & a_m \end{bmatrix}^T = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

We can represent a matrix $\mathbf{X}$ as a partitioned matrix whose generic block is the $1 \times n$ row vector $\mathbf{x}_i^T = [x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{in}]$, which contains the profile of the $i$-th row unit,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_m^T \end{bmatrix}$$

Alternatively, we can partition as

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_j, ..., \mathbf{x}_n],$$

where $\mathbf{x}_j$ is the $m \times 1$ column vector referring to the $j$-th variable or attribute.

## 2.3. **Transpose.**

The **transpose** of an $m \times n$ matrix $\boldsymbol{A}$ yields an $n \times m$ matrix that interchanges the rows and columns of $\boldsymbol{A}$.

$$\mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{i1} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{i2} & \cdots & a_{m2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{1j} & a_{2j} & \cdots & a_{ij} & \cdots & a_{mj} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{in} & \cdots & a_{mn} \end{bmatrix}$$

The transpose has the property that $(\mathbf{A}^T)^T = \mathbf{A}$

**Example.**

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 4 \\ 1 & 2 & -1 \\ 3 & -2 & 5 \\ -2 & 4 & 1 \end{bmatrix}, \qquad \mathbf{A}^T = \begin{bmatrix} 2 & 1 & 3 & -2 \\ 3 & 2 & -2 & 4 \\ 4 & -1 & 5 & 1 \end{bmatrix}$$

## 2.4. **Addition and scalar multiplication.**

**Scalar Multiplication**. $\alpha\mathbf{a} = \{\alpha a_{ij}\}$.

**Matrix addition**. If $\boldsymbol{A}$ is an $m \times n$ matrix and $\boldsymbol{B}$ an $m \times n$ matrix, then

$$\boldsymbol{A} + \boldsymbol{B} = \{a_{ij} + b_{ij}\}.$$

This can only be performed if $\boldsymbol{A}$ and $\boldsymbol{B}$ have the exact same dimensions.

Note that $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$.

**Example.**

$$\begin{bmatrix} 2 & 3 \\ 3 & -2 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 2+1 & 3+2 \\ 3+3 & -2+4 \end{bmatrix} = \begin{bmatrix} 3 & 5 \\ 6 & 2 \end{bmatrix}$$

## 2.5. **Matrix-vector multiplication.**

The product of an $m \times n$ matrix $\boldsymbol{A}$ with an $n$-vector $\boldsymbol{b}$ is an $m$-vector $\boldsymbol{c}$ with element $i$ equal to the inner product of the row $i$ of $\boldsymbol{A}$ with $\boldsymbol{b}$.

$$c_i = \boldsymbol{a}_i^T \boldsymbol{b} = \sum_{j=1}^{n} a_{ij} b_j,$$

where $\boldsymbol{a}_i^T$ denotes $i$-th row of $\boldsymbol{A}$.

**Example.**

$$\begin{bmatrix} 1 & 4 \\ 7 & -3 \\ 2 & -5 \end{bmatrix} \times \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \times 2 + 4 \times 1 \\ 7 \times 2 - 3 \times 1 \\ 2 \times 2 - 5 \times 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 11 \\ -1 \end{bmatrix}$$

2.6. **Matrix-matrix multiplication.**

The product of an $m \times p$ matrix $\boldsymbol{A}$ with an $p \times n$ matrix $\boldsymbol{B}$ is an $m \times n$ matrix $\boldsymbol{C}$ with element $ij$ equal to the inner product of the row $i$ of $\boldsymbol{A}$ with column $j$ of $\boldsymbol{B}$

$$c_{ij} = \boldsymbol{a}_i^T \boldsymbol{b}_j = \sum_{k=1}^{p} a_{ik} b_{kj}$$

where $\boldsymbol{a}_i^T$ denotes $i$-th row of $\boldsymbol{A}$ and $\boldsymbol{b}_j$ denotes $j$-th column of $\boldsymbol{B}$.

The matrix partitions that we use in the multiplication are

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_i^T \\ \vdots \\ \mathbf{a}_m^T \end{bmatrix}, \quad \boldsymbol{B} = [\mathbf{b}_1, ..., \mathbf{b}_j, ..., \mathbf{b}_n]$$

The multiplication ($\boldsymbol{AB}$) is only defined when the column dimension of $\boldsymbol{A}$ ($m \times p$ matrix) equals the row dimension$\boldsymbol{B}$ ($p \times n$ matrix).

**Example.**

$$\begin{bmatrix} 2 & 3 \\ 3 & -2 \end{bmatrix} \times \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 2 \times 1 + 3 \times 3 & 2 \times 2 + 3 \times 4 \\ 3 \times 1 - 2 \times 3 & 3 \times 2 - 2 \times 4 \end{bmatrix} = \begin{bmatrix} 11 & 16 \\ -3 & -2 \end{bmatrix}$$

**Properties of matrix multiplication.**

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$
$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$$
$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$$
$$(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$$

Unlike in scalar multiplication, the order of multiplication matters for matrices: in general, $\mathbf{AB}\neq\mathbf{BA}$. Moreover, remember that if $m \neq n$, $\mathbf{BA}$ is not even defined.

**Example.**

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 5 & -1 \\ 3 & 2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 2 & -1 \\ 3 & 6 \end{bmatrix}, \quad \boldsymbol{\iota}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 2 & -1 \\ 7 & -11 \\ 12 & 9 \end{bmatrix}, \quad \boldsymbol{\iota}^T\mathbf{C} = \begin{bmatrix} 21 & -3 \end{bmatrix}.$$

$\mathbf{BA}$ is not defined.

**Vector outer product.** If $\mathbf{a}$ is an $m$-vector and $\mathbf{b}$ is an $n$- vector, the outer product $\mathbf{ab}^T$ is the $m \times n$ matrix

$$\mathbf{ab}^T = \begin{bmatrix} a_1b_1 & a_1b_2 & \dots & a_1b_n \\ a_2b_1 & a_2b_2 & \dots & a_2b_n \\ \vdots & \vdots & & \vdots \\ a_mb_1 & a_mb_2 & \dots & a_mb_n \end{bmatrix}$$

## 2.7. **Square matrices.**

A **square matrix** has the same number or rows and columns $m = n$.

**Symmetric matrix.** A square matrix $\mathbf{A}$ is symmetric if $\mathbf{A}^T = \mathbf{A}$.

**Quadratic form**. Let $\mathbf{A}$ be an $n$ dimensional square matrix and $\mathbf{x}$ an $n \times 1$ vector. The scalar $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is called a quadratic form.

A symmetric matrix $\mathbf{A}$ with the property that $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for any vector $\mathbf{x}$ is said to be **positive definite**.

## 2.8. **Identity and diagonal matrices.**

The diagonal elements of a matrix are the elements $a_{ij}$ such that $i = j$ (same row and column index).

An **identity matrix** of order $n$ is a matrix with all diagonal elements equal to one ($a_{ii} = 1$ for $i = 1, \ldots, n$), and all non-diagonal elements equal to zero, that is

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} = \text{diag}(1, ..., 1)$$

For example,

$$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

**Properties.**

Let $\boldsymbol{A}$ be an $m \times n$ matrix.

$$\boldsymbol{I}_n^2 = \boldsymbol{I}_n$$
$$\boldsymbol{I}_m\boldsymbol{A} = \boldsymbol{A}$$

$$\boldsymbol{A}\boldsymbol{I}_n = \boldsymbol{A}$$

**Diagonal matrix.** A diagonal matrix is a square matrix with zeros in all the non-diagonal positions.

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 & 0 \\ 0 & d_2 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & d_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & d_n \end{bmatrix} = \text{diag}(d_1, ..., d_n)$$

Let $\boldsymbol{D}$ be an $n \times n$ diagonal matrix and $\boldsymbol{A}$ an $n \times p$ matrix. The operation $\boldsymbol{D}\boldsymbol{A}$ multiplies each row $i$ of $\boldsymbol{A}$ by the diagonal element $d_i$ of $\boldsymbol{D}$.

## 2.9. **Systems of Linear Equations.**

Consider a system of $m$ linear equations with $n$ variables.

$$\begin{gathered} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_2 \end{gathered}$$

This system has a compact representation in matrix notation

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

where

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{21} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}, \ \boldsymbol{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}, \ \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}.$$

The study of linear systems is a fundamental part of linear algebra, which allows us to determine whether a system has an unique solution, infinitely many solutions, or no solutions, and to obtain a solution if one exists.

**Example.**

We can write the system

$$\begin{cases} 2x_1 + 2x_2 + x_3 &= 9 \\ 2x_1 - x_2 + 2x_3 &= 6 \\ x_1 - x_2 + 2x_3 &= 5 \end{cases}$$

as

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

where

$$\boldsymbol{A} = \begin{bmatrix} 2 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & -1 & 2 \end{bmatrix}, \ \boldsymbol{b} = \begin{bmatrix} 9 \\ 6 \\ 5 \end{bmatrix}.$$

The unique solution is $\boldsymbol{x} = (1, 2, 3)$.

## 2.10. **Matrix inverse.**

An $n \times n$ matrix $\boldsymbol{A}$ is **invertible** is there exists a matrix $\boldsymbol{B}$ such that

$$\boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}_n$$

If that is the case then we call $\boldsymbol{B}$ the **inverse** of $\boldsymbol{A}$, and use the notation $\boldsymbol{A}^{-1}$.

There are several methods for calculating a matrix inverse, but we will leave the details in the background. It is often the case in practice that the we do not actually need to explictly compute the matrix inverse to evaluate expressions in which it appears (for example in the formula for the OLS).

**Properties.**

$$(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$$
$$(\alpha\boldsymbol{A})^{-1} = (1/\alpha)\boldsymbol{A}^{-1}$$
$$(\boldsymbol{A}^{T})^{-1} = (\boldsymbol{A}^{-1})^{T}$$
$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$$

### 2.11. **Trace**[†]**.**

The **trace** of a square matrix is the sum of its diagonal elements. If $\mathbf{A}$ is $n \times n$,

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

**Properties**.

$$\operatorname{tr}(\alpha\mathbf{A}) = \alpha \operatorname{tr}(\mathbf{A})$$
$$\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B})$$
$$\operatorname{tr}(\mathbf{A}^{T}) = \operatorname{tr}(\mathbf{A})$$
$$\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$$

## 3. Differentiation

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a function. The **gradient** of $f$ is the vector of partial derivatives of the function with respect to each of its arguments.

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

We also use the notation:

$$\frac{d\, f(\boldsymbol{x})}{d\boldsymbol{x}} = \left( \frac{\partial f(\boldsymbol{x})}{\partial x_1}, \frac{\partial f(\boldsymbol{x})}{\partial x_2}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_n} \right)$$

There are several convenient rules for differentiating linear algebra operations with respect to vectors. The two following rules appear in the derivation of the least squares estimates of a linear regression.

Let $\boldsymbol{x}$ and $\boldsymbol{a}$ be $n-$vectors and $\boldsymbol{A}$ a matrix with column dimension $n$. Then,

$$\frac{d(\boldsymbol{x}'\boldsymbol{a})}{d\boldsymbol{x}} = \boldsymbol{a}$$

$$\frac{d(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})}{d\boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{x}$$

The `Matrix Algebra Cookbook` found online contains a comprehensive catalog.

## 4. Random vectors

A **random vector** or **multivariate random variable** is a vector with entries that are scalar-valued random variables.

Let $X = (X_1\, X_2\, \ldots\, X_n)$ be a random vector. The **mean vector**, or **expected value** of $X$ is a $n$-vector over $\mathbb{R}$ defined as

$$E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix}$$

Let $a$ and $b$ be non-random scalars and $Y$ be another random vector with dimension $n$. Then

$$E(aX + bY) = aE(X) + bE(Y),$$

which follows from the linearity of expectations.

For a non-random $n-$vector $\boldsymbol{a}$,

$$E(\boldsymbol{a}^T X) = \boldsymbol{a}^T E(X).$$

Let $\boldsymbol{A}$ be a non-random matrix with $n$ columns.

$$E(\boldsymbol{A}X) = \boldsymbol{A}E(X)$$

We define the **variance** of the random vector as the square matrix

$$\mathrm{Var}(X) = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & \mathrm{Var}(X_n) \end{bmatrix}$$

$\mathrm{Var}(X)$ is also known as the variance-covariance or covariance matrix of $X$.

$$\mathrm{Var}(X) = E(XX^T) - E(X)E(X)^T$$

Let $\boldsymbol{a}$ be a non-random vector and $b$ a scalar. Then,

$$\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)$$

For a non-random $n-$vector $\boldsymbol{a}$,

$$\mathrm{Var}(\boldsymbol{a}^T X) = \boldsymbol{a}^T\mathrm{Var}(X)\boldsymbol{a}.$$

For a non-random matrix $\boldsymbol{A}$ with $n$ columns,

$$\mathrm{Var}(\boldsymbol{A}X) = \boldsymbol{A}\mathrm{Var}(X)\boldsymbol{A}^T.$$

## 5. Application: linear regression and least squares

### 5.1. **Multiple Linear Regression (MLR) model.**

The classical MLR model is characterised by the following set of assumptions.

1. Linearity: if $X = \boldsymbol{x}$, then
$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$
for some population parameters $\beta_0, \beta_1, \ldots, \beta_p$ and a random error $\varepsilon$.
2. The conditional mean of $\varepsilon$ given $X$ is zero, $E(\varepsilon|X) = 0$.
3. Constant error variance: $\mathrm{Var}(\varepsilon|X) = \sigma^2$.
4. Independence: the observations are independent.
5. The distribution of $X_1, \ldots, X_p$ is arbitrary.
6. There is no perfect multicollinearity (no column of $\boldsymbol{X}$ is a linear combination of other columns).

The model equation for an observed indexed by $i$ is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

We can therefore write the model of a sample of $n$ observations as

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{np} + \varepsilon_n \end{aligned}$$

Compact matrix notation:

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Assumptions 2 and 4 imply that

$$\mathrm{Var}(\varepsilon) = \sigma^2 \boldsymbol{I}_n$$

## 5.2. **Least squares.**

Let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be a sample. The **ordinary least squares** (OLS) method obtains the coefficient values that minimise the residual sum of squares (RSS):

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

The partial derivatives of the RSS with respect to the coefficients are

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)$$

$$\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \beta_1} = -2\sum_{i=1}^{n} x_{i1}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)$$
$$\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \beta_2} = -2\sum_{i=1}^{n} x_{i2}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)$$
$$\vdots$$
$$\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \beta_p} = -2\sum_{i=1}^{n} x_{ip}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)$$

Note that each partial derivative $j$ contains a sum that is the inner product of the $j$-th column of $\boldsymbol{X}$ with the vector of residuals $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. We can therefore write the above equations using the compact notation

$$\bigtriangledown \text{RSS}(\boldsymbol{\beta}) = -2\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

We obtain the first order condition by setting the gradient to zero,

$$\frac{d(\text{RSS}(\boldsymbol{\beta}))}{d\boldsymbol{\beta}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}.$$

The least squares estimate $\widehat{\boldsymbol{\beta}}$ therefore satisfies the system of linear equations:

$$\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}$$

Note that $\boldsymbol{X}^T\boldsymbol{X}$ is a $(p+1) \times (p+1)$ matrix and $\boldsymbol{X}^T\boldsymbol{y}$ is a $(p+1)$-vector, such that this expression has the form of Section 2.9.

If $(\boldsymbol{X}^T\boldsymbol{X})$ is invertible, which is the case if Assumption 6 of no perfect multicollinearity is satisfied, left multiplication with $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ gives the unique solution

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

**Solution using vector differentiation rules.**

$$\begin{aligned}\mathrm{RSS}(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\end{aligned}$$

The gradient is

$$\begin{aligned}\frac{d(\mathrm{RSS}(\boldsymbol{\beta}))}{d\boldsymbol{\beta}} = \frac{d(\boldsymbol{y}^T\boldsymbol{y})}{d\boldsymbol{\beta}} - \frac{d(2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y})}{d\boldsymbol{\beta}} + \frac{d(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta})}{d\boldsymbol{\beta}} = \\ = \mathbf{0} - 2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\end{aligned}$$

The first order condition is therefore

$$\frac{d(\mathrm{RSS}(\boldsymbol{\beta}))}{d\boldsymbol{\beta}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \mathbf{0}.$$

Therefore, as above

$$\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y},$$

leading to

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

## 5.3. **Sampling properties.**

We first obtain the following convenient representation of the estimator.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}
\end{aligned}$$

Below, all the results are conditional on the predictor values in the $\boldsymbol{X}$ matrix. We omit this conditioning from the notation for simplicity.

**Expected value.**

$$\begin{aligned}
E(\widehat{\boldsymbol{\beta}}) &= E(\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}) \\
&= \boldsymbol{\beta} + E\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}\right] \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{\varepsilon}) \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{0} \\
&= \boldsymbol{\beta}
\end{aligned}$$

The least squares estimator is unbiased under the model assumptions.

**Variance.**

$$
\begin{aligned}
\mathrm{Var}(\widehat{\boldsymbol{\beta}}) &= \mathrm{Var}(\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}) \\
&= \mathrm{Var}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}) \\
&= E((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}
\end{aligned}
$$

## References


Boyd, S. and L. Vandenberghe (2016). Vectors, matrices, and least squares. *Available: stanford.edu/class/ee103/mma.pdf*.

Klein, P. N. (2013). *Coding the matrix: Linear algebra through applications to computer science*. Newtonian Press.