Statistics II
Professor: Marcelo J. Moreira
TA: Luan Borelli

Review
Linear Algebra and Multivariate Statistics
July 07, 2024

This note provides an informal, non-comprehensive, and still-evolving introduction to linear algebra, matrix algebra, and multivariate statistics. The order of exposition may not be optimal at this stage. Consider it a collection of basic definitions, results, and comments that can help you warm up for the Statistics II course. I highly recommend becoming comfortable with the notation, concepts, and main results presented in this note as soon as possible, so you can better enjoy the course. Have fun.

1 Row and Column Vectors

A *vector* v is an element of a vector space. In analysis we rarely speak of *column* or *row* vectors, as such terminology is generally unnecessary. However, in the context of matrix algebra, one might need to perform matrix operations involving vectors. In such cases, it becomes important to define precisely what a vector is in the language of matrices.

Definition 1 (Row vector)**.** A n-dimensional *row vector* $\boldsymbol{v}$ is an $1 \times n$ matrix

$$\boldsymbol{v} = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}.$$

We denote $(v_1, v_2, \ldots, v_n) \equiv \boldsymbol{v}$.

Definition 2 (Column vector)**.** A n-dimensional *column vector* $\boldsymbol{v}$ is an $n \times 1$ matrix

$$\boldsymbol{v} = (v_1, v_2, \cdots, v_n)' = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix},$$

where $(\cdot)'$ denotes the usual transposition operation.

In matrix algebra, vectors are generally treated as *column* vectors. One might wonder why this convention is adopted. A plausible explanation is that the column convention has the appealing property that if $\boldsymbol{v}$ is a vector and $\boldsymbol{M}$ is a matrix representing a linear transformation, the product $\boldsymbol{M}\boldsymbol{v}$, computed using the usual rules of matrix multiplication, is another vector (specifically, a column vector) representing the image of $\boldsymbol{v}$ under that transformation.

2 Partitions and Conformable Partitioning

Definition 3 (Partition)**.** Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, where $\mathbb{R}^{m \times n}$ denotes the space of $m \times n$ real matrices.[1] A *partitioning* of $\boldsymbol{A}$ is a representation of $\boldsymbol{A}$ in the form of

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1q} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}_{p1} & \boldsymbol{A}_{p2} & \cdots & \boldsymbol{A}_{pq} \end{bmatrix},$$

where $\boldsymbol{A}_{ij} \in \mathbb{R}^{m_i \times n_j}$ are contiguous submatrices, $\sum_{i=1}^{p} m_i = m$, and $\sum_{j=1}^{q} n_j = n$. The elements $\boldsymbol{A}_{ij}$ of the partition are called *blocks*.

Let's play around a bit with partitions. We all know how to multiply matrices. Let $\boldsymbol{A} \in \mathbb{R}^{4 \times 2}$ and $\boldsymbol{B} \in \mathbb{R}^{2 \times 2}$. As the number of columns of $\boldsymbol{A}$ equals the number of rows of $\boldsymbol{B}$, the matrix multiplication $\boldsymbol{AB}$ is well-defined:

$$\boldsymbol{AB} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} \\ a_{41}b_{11} + a_{42}b_{21} & a_{41}b_{12} + a_{42}b_{22} \end{bmatrix}. \tag{1}$$

Now, define $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ as the upper and bottom 2×2 blocks of $\boldsymbol{A}$, respectively; that is,

$$\boldsymbol{A}_1 \equiv \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \text{and} \quad \boldsymbol{A}_2 \equiv \begin{bmatrix} a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix}.$$

This gives us the following partitioning of $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix}.$$

Now, $\boldsymbol{A}$ is structured as a 2×1 block matrix consisting of two 2×2 blocks, $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$. This partitioning appears to be reasonable — just an alternative way of expressing $\boldsymbol{A}$. Therefore, the product $\boldsymbol{AB}$ under this partitioning should still be well-defined... Correct?

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \ldots \boldsymbol{?}$$

Actually, it is not! Under this partitioning, $\boldsymbol{A}$ becomes a 2×1 block matrix, while $\boldsymbol{B}$ remains a 2×2 real scalar matrix. Consequently, the number of columns in $\boldsymbol{A}$ no longer matches the number of rows in $\boldsymbol{B}$. As a result, the product $\boldsymbol{AB}$ doesn't make sense; it becomes

[1] Let S be a set. More generally, one could define the $m \times n$ *matrix space over* S as the the set of all $m \times n$ matrices over S and denote it by $S^{m \times n}$.

unconformable. Loosely speaking, two matrices are said to be *conformable* with respect to a given operation if they possess the necessary traits for that operation to be well-defined.

What if we also partition $\boldsymbol{B}$ by letting $\boldsymbol{b}_1 = (b_{11}, b_{21})'$ and $\boldsymbol{b}_2 = (b_{12}, b_{22})'$ be the first and second columns of $\boldsymbol{B}$, so that $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{b}_1 & \boldsymbol{b}_2\end{bmatrix}$? Then we have

$$\boldsymbol{AB} = \begin{bmatrix}\boldsymbol{A}_1 \\ \boldsymbol{A}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{b}_1 & \boldsymbol{b}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{A}_1\boldsymbol{b}_1 & \boldsymbol{A}_1\boldsymbol{b}_2 \\ \boldsymbol{A}_2\boldsymbol{b}_1 & \boldsymbol{A}_2\boldsymbol{b}_2\end{bmatrix} = \begin{bmatrix}a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} \\ a_{41}b_{11} + a_{42}b_{21} & a_{41}b_{12} + a_{42}b_{22}\end{bmatrix}.$$

Under this new additional partitioning of $\boldsymbol{B}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ become *partitioned conformably*, making the product operation $\boldsymbol{AB}$ conformable again. The resulting product is equivalent to the original non-partitioned product of $\boldsymbol{A}$ and $\boldsymbol{B}$ presented in (1).

Definition 4 (Product-conformable partitioning)**.** Two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to be *partitioned conformably* with respect to the product $\boldsymbol{AB}$ when $\boldsymbol{A}$ and $\boldsymbol{B}$ are partitioned into blocks and the multiplication $\boldsymbol{AB}$ can be carried out treating the blocks as if they were scalars, but keeping the order, and all products and sums of blocks involved are conformable.

Needless to say, conformable partitioning is necessary not only for product operations but also for sum and subtraction operations.

In summary, any matrix can be interpreted as a block matrix in one or more ways, with each interpretation defined by how its rows and columns are partitioned. Partitioning is very common in econometrics and can be extremely useful in matrix algebra, often greatly simplifying algebraic derivations. However, partitioning is not arbitrary; it requires conformable partitions between two (or more) matrices to ensure that all the submatrix operations involved are well-defined.

Example 1. Consider the linear model

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i, \quad \forall i = 1, 2, \ldots, n. \tag{2}$$

By stacking y_i for all $i = 1, 2, \ldots, n$ we can write

$$\begin{aligned}\begin{bmatrix}y_1 \\ y_2 \\ \vdots \\ y_n\end{bmatrix} &= \begin{bmatrix}1 \\ 1 \\ \vdots \\ 1\end{bmatrix}\beta_1 + \begin{bmatrix}x_{21} \\ x_{22} \\ \vdots \\ x_{2n}\end{bmatrix}\beta_2 + \cdots + \begin{bmatrix}x_{k1} \\ x_{k2} \\ \vdots \\ x_{kn}\end{bmatrix}\beta_k + \begin{bmatrix}u_1 \\ u_2 \\ \vdots \\ u_n\end{bmatrix} \\ &= \begin{bmatrix}1 & x_{21} & \cdots & x_{k1} \\ 1 & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & \cdots & x_{kn}\end{bmatrix}\begin{bmatrix}\beta_1 \\ \beta_2 \\ \vdots \\ \beta_k\end{bmatrix} + \begin{bmatrix}u_1 \\ u_2 \\ \vdots \\ u_n\end{bmatrix},\end{aligned}$$

which gives us the matrix representation

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{u}. \tag{3}$$

A common partitioning of $\boldsymbol{X}$ is obtained by defining $\boldsymbol{x}_j = (1, x_{2j}, x_{3j}, \ldots, x_{kj})'$ for all $i = 1, 2, \ldots, n$ and writing

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix}.$$

Notice that, for each j, $\boldsymbol{x}_j'$ is a $1 \times k$ row vector and $\boldsymbol{\beta}$ a $k \times 1$ column vector, so that

$$\boldsymbol{X\beta} = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{x}_1'\boldsymbol{\beta} \\ \boldsymbol{x}_2'\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{x}_n'\boldsymbol{\beta} \end{bmatrix}.$$

Then we have

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1'\boldsymbol{\beta} \\ \boldsymbol{x}_2'\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{x}_n'\boldsymbol{\beta} \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix},$$

or

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \quad \forall i = 1, \ldots, n. \tag{4}$$

Equations (2), (3), and (4) give three alternative representations for the same linear model: a full scalar representation, a full matrix representation, and a vector representation. The three representations are useful in econometrics, and it is important to be familiar with all of them, being able to transition naturally from one to another. △

3 Properties of Transposes and Orthogonal Matrices

Please recall the following basic properties of transposes.

1. $(\boldsymbol{A}')' = \boldsymbol{A}$.
2. $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$.
3. $(\boldsymbol{A} + \boldsymbol{B})' = \boldsymbol{A}' + \boldsymbol{B}'$.
4. $(\boldsymbol{A}^{-1})' = (\boldsymbol{A}')^{-1}$.
5. $(c\boldsymbol{A})' = c\boldsymbol{A}'$, where c is a scalar.
6. $\det(\boldsymbol{A}') = \det(\boldsymbol{A})$.

7. For a partitioning $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1q} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}_{p1} & \boldsymbol{A}_{p2} & \cdots & \boldsymbol{A}_{pq} \end{bmatrix}$, $\boldsymbol{A}' = \begin{bmatrix} \boldsymbol{A}'_{11} & \boldsymbol{A}'_{21} & \cdots & \boldsymbol{A}'_{p1} \\ \boldsymbol{A}'_{12} & \boldsymbol{A}'_{22} & \cdots & \boldsymbol{A}'_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}'_{1q} & \boldsymbol{A}'_{2q} & \cdots & \boldsymbol{A}'_{pq} \end{bmatrix}$.

Definition 5 (Orthogonal matrix)**.** We say that a $m \times n$ matrix $\boldsymbol{A}$ with $m \geq n$ is *orthogonal* (or *orthonormal*) if $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Please note that $\boldsymbol{A}$ being orthogonal does *not* imply that $\boldsymbol{A}' = \boldsymbol{A}^{-1}$ or $\boldsymbol{A} = (\boldsymbol{A}')^{-1}$. This property holds true only when $\boldsymbol{A}$ is a square matrix (i.e., when $m = n$). It doesn't make sense to discuss inverses of non-square matrices.[2]

4 Quadratic Forms and Positive (Semi)Definiteness

Definition 6 (Quadratic form)**.** A *quadratic form* is a multivariate polynomial $q(\boldsymbol{x})$ with terms all of degree two,

$$q(x_1, x_2, \dots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

Observe that any quadratic form can be written, using matrix notation, in the form $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Indeed, letting $\boldsymbol{x} = (x_1, x_2, \dots, x_n)'$ and $\boldsymbol{A} = \begin{bmatrix} a_{ij} \end{bmatrix}$,

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{5}$$

$$= \begin{bmatrix} \sum_{i=1}^n a_{i1} x_i & \sum_{i=1}^n a_{i2} x_i & \cdots & \sum_{i=1}^n a_{in} x_i \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{6}$$

$$= \sum_{j=1}^{n} x_j \sum_{i=1}^{n} a_{ij} x_i = \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} x_i x_j = q(x_1, x_2, \dots, x_n). \tag{7}$$

Also note that general quadratic forms encompass several interesting particular cases, depending on the form of $\boldsymbol{A}$. For instance, if $\boldsymbol{A}$ is diagonal, then $a_{ij} = 0$ for all $i \neq j$; hence, $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^n a_{ii} x_i^2$. If $\boldsymbol{A} = \boldsymbol{I}_n$, then $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{x} = \sum_{i=1}^n x_i^2$; that is, the *dot product* of $\boldsymbol{x}$ with itself.

Definition 7 (Dot product)**.** Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$. The *dot product* of $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as

$$\boldsymbol{a}'\boldsymbol{b} = a_1 b_1 + a_2 b_2 + \cdots + a_k b_k = \sum_{i=1}^{n} a_i b_i.$$

[2] It does make sense, however, to discuss Moore-Penrose inverses — also known as *pseudo-inverses*.

Quadratic forms relate to definiteness of matrices.

Definition 8. An $n \times n$ symmetric matrix $\boldsymbol{A}$ is said to be *positive definite* (PD) if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n \backslash \{0\}$, and *positive semidefinite* (PSD) if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

We often denote "$\boldsymbol{A}$ is PD" by $\boldsymbol{A} \succ 0$, and "$\boldsymbol{A}$ is PSD" by $\boldsymbol{A} \succeq 0$. Sometimes, $>$ and $\geq$ are used instead. It's important to note that this is just a notation. $\boldsymbol{A} > 0$ does *not* imply that all entries of $\boldsymbol{A}$ are strictly positive or anything of that sort.

Loosely speaking, positive definiteness is an important concept in Econometrics as it serves as a kind of generalization of scalar positivity to matrices. For a scalar random variable with variance σ^2, we require the variance to be nonnegative; i.e., $\sigma^2 \geq 0$. Similarly, for a vector random variable, we aim to ensure that its variance-covariance matrix $\boldsymbol{\Sigma}$ is "nonnegative" in a suitable sense. This is achieved by requiring $\boldsymbol{\Sigma}$ to be nonnegative in the PSD sense, denoted as $\boldsymbol{\Sigma} \succeq 0$.

Moreover, the notion of positive definiteness allows us to establish a partial order over symmetric matrices, enabling comparisons between two such matrices. This partial order is established by defining the binary operation $\succeq$ as $\boldsymbol{A} \succeq \boldsymbol{B} \iff \boldsymbol{A} - \boldsymbol{B}$ is PSD. This order is also known as the *Loewner order*. When $\boldsymbol{A}$ and $\boldsymbol{B}$ are scalars, the Loewner order reduces to the usual ordering in $\mathbb{R}$.

Establishing an order is important, especially in the context of variance analysis. For example, when comparing the variances of two random variables, such as two different estimators $\hat{\theta}$ and $\tilde{\theta}$, it is natural to ask which estimator has the smaller variance. If $\hat{\theta}$ and $\tilde{\theta}$ are real scalar random variables, their variances $\text{var}(\hat{\theta})$ and $\text{var}(\tilde{\theta})$ are also scalars, making comparison straightforward using the usual ordering of $\mathbb{R}$. However, when $\hat{\theta}$ and $\tilde{\theta}$ are vector random variables, their variances $\text{var}(\hat{\theta})$ and $\text{var}(\tilde{\theta})$ are variance-covariance matrices. Determining which estimator has the smallest variance becomes less straightforward. Econometricians typically state that $\hat{\theta}$ has a smaller variance than $\tilde{\theta}$ if $\text{var}(\tilde{\theta}) - \text{var}(\hat{\theta}) \succ 0$, indicating that the difference between their variance-covariance matrices is positive definite.

5 Row/Column Spaces and Rank

Definition 9 (Column space)**.** Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. Let $v_1, \ldots, v_n \in \mathbb{R}^{m \times 1}$ be the column vectors of $\boldsymbol{A}$. The *column space* of $\boldsymbol{A}$ is the set of all possible linear combinations of $v_1, \ldots, v_n$.

In other words, the column space of $\boldsymbol{A}$ is the space spanned by the column vectors of $\boldsymbol{A}$:

$$\text{colsp}(\boldsymbol{A}) = \text{span}(v_1, \ldots, v_n).$$

Definition 10 (Row space)**.** Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. Let $r_1, \ldots, r_m \in \mathbb{R}^{1 \times n}$ be the column vectors of $\boldsymbol{A}$. The *row space* of $\boldsymbol{A}$ is the set of all possible linear combinations of $r_1, \ldots, r_m$.

In other words, the row space of $\boldsymbol{A}$ is the space spanned by the row vectors of $\boldsymbol{A}$:

$$\text{rowsp}(\boldsymbol{A}) = \text{span}(r_1, \ldots, r_m).$$

Definition 11 (Basis)**.** A basis B of a vector space V is a linearly independent subset of V that spans V.

In other words, a basis is a linearly independent spanning set. It's important to note that the columns (rows) of $\boldsymbol{A}$ span the column (row) space of $\boldsymbol{A}$, but they do not necessarily form a basis for this space, since $\boldsymbol{A}$ may contain linearly dependent columns (rows). However, a maximal linearly independent subset of the column (row) vectors does provide a basis for the column (row) space.

Definition 12 (Column and row rank)**.** The column (row) rank of $\boldsymbol{A}$ is the dimension of the column (row) space of $\boldsymbol{A}$.

Please recall that the dimension of the column (row) space of $\boldsymbol{A}$ is the dimension of its basis.

Theorem 1. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. The row rank of $\boldsymbol{A}$ is equal to the column rank of $\boldsymbol{A}$.*

Due to Theorem 1, we refer to the column (or row) rank of $\boldsymbol{A}$ simply as the *rank* of $\boldsymbol{A}$.

Definition 13 (Full rank)**.** Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. We say that $\boldsymbol{A}$ has *full rank* if $\text{rank}(\boldsymbol{A}) = \min\{m, n\}$.

Why $\min\{m, n\}$? Suppose, on the contrary, that $m < n$ and $\text{rank}(\boldsymbol{A}) = n$. Then $\text{rank}(\boldsymbol{A}) = \dim(\text{rowsp}(\boldsymbol{A})) \leq m < n = \text{rank}(\boldsymbol{A})$, a contradiction. The first inequality follows form the fact that if the row vectors of form a basis for the row space of $\boldsymbol{A}$, then $\dim(\text{rowsp}(\boldsymbol{A})) = m$; and, if not, then at least two of the m row vectors of $\boldsymbol{A}$ are linearly dependent, so that $\dim(\text{rowsp}(\boldsymbol{A})) < m$.

Theorem 2 (Full rank and invertibility)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. $\boldsymbol{A}$ is invertible (or nonsingular) if and only if $\boldsymbol{A}$ is full rank.*

6 Properties of Inverses

Please recall the following basic properties of inverses.

1. $(A^{-1})^{-1} = A$
2. $(cA)^{-1} = \frac{1}{c}A^{-1}$, where c is a scalar.
3. $(A')^{-1} = (A^{-1})'$
4. $(AB)^{-1} = B^{-1}A^{-1}$
5. $(A_1A_2\cdots A_{n-1}A_n)^{-1} = A_n^{-1}A_{n-1}^{-1}\cdots A_2^{-1}A_1^{-1}$.
6. For a conformable partitioning $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{A}_3 & \boldsymbol{A}_4 \end{bmatrix}$,

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}_1^{-1} + \boldsymbol{A}_1^{-1}\boldsymbol{A}_2(\boldsymbol{A}_4 - \boldsymbol{A}_3\boldsymbol{A}_1^{-1}\boldsymbol{A}_2)^{-1}\boldsymbol{A}_3\boldsymbol{A}_1^{-1} & -\boldsymbol{A}_1^{-1}\boldsymbol{A}_2(\boldsymbol{A}_4 - \boldsymbol{A}_3\boldsymbol{A}_1^{-1}\boldsymbol{A}_2)^{-1} \\ -(\boldsymbol{A}_4 - \boldsymbol{A}_3\boldsymbol{A}_1^{-1}\boldsymbol{A}_2)^{-1}\boldsymbol{A}_3\boldsymbol{A}_1^{-1} & (\boldsymbol{A}_4 - \boldsymbol{A}_3\boldsymbol{A}_1^{-1}\boldsymbol{A}_2)^{-1} \end{bmatrix},$$

provided $\boldsymbol{A}_1$ and the Schur complement of $\boldsymbol{A}_1$ in $\boldsymbol{A}$, $\boldsymbol{A}_4 - \boldsymbol{A}_3\boldsymbol{A}_1^{-1}\boldsymbol{A}_2$, are invertible.[3]

I stress, again, that it does not make sense to discuss inverses of non-square matrices. For example, you should be familiar with the OLS estimator in its matrix form:

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

One might naively think that this formula could be simplified by applying the fourth property and then writing

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}^{-1}(\boldsymbol{X}')^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}^{-1}\boldsymbol{Y}.$$

Of course, this doesn't make any sense. Since in general, the design matrix $\boldsymbol{X}$ has dimensions $n \times k$, it is not a square matrix, and therefore, "$\boldsymbol{X}^{-1}$" is not a well-defined object.

7 Properties of Determinants

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Please recall the following basic properties of determinants.

1. $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$
2. $\det(c\boldsymbol{A}) = c^n \det(\boldsymbol{A})$, where c is a scalar.
3. $\det(\boldsymbol{AB}) = \det(\boldsymbol{BA}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$, for any $\boldsymbol{B} \in \mathbb{R}^{n\times n}$.
4. $\det(\boldsymbol{A}^{-1}) = (\det(\boldsymbol{A}))^{-1}$
5. For a conformable partitioning $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{A}_3 & \boldsymbol{A}_4 \end{bmatrix}$ of $\boldsymbol{A}$,

$$\det \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{0} & \boldsymbol{A}_4 \end{bmatrix} = \det \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{0} \\ \boldsymbol{A}_3 & \boldsymbol{A}_4 \end{bmatrix} = \det \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_4 \end{bmatrix} = \det(\boldsymbol{A}_1\boldsymbol{A}_4) = \det(\boldsymbol{A}_1)\det(\boldsymbol{A}_4).$$

6. If $\boldsymbol{A}$ is triangular (or diagonal, in particular), then $\det(\boldsymbol{A}) = \prod_{i=1}^n a_{ii}$.
7. $\det(\boldsymbol{A}) = \prod_{i=1}^n \lambda_i$, where λ_i, $i = 1, \ldots, n$, are the eigenvalues of $\boldsymbol{A}$.[4]

Just as with inversions, please note that it doesn't make sense to talk about determinants of non-square matrices.

[3] Of course, you don't need to memorize this property. However, it is interesting to observe how things simplify when $\boldsymbol{A}_2$ and/or $\boldsymbol{A}_3$ are zero matrices; specifically, when $\boldsymbol{A}$ is a block (upper or lower) triangular or block diagonal matrix.

[4] Recall that the *eigenvalues* of $\boldsymbol{A}$ are the roots of the polynomial $\det(\lambda\boldsymbol{I}_n - \boldsymbol{A}) = 0$.

8 Properties of Traces

Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. We define the trace operator tr as $\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}$. Please recall the following basic properties of traces.

1. $\operatorname{tr}(\boldsymbol{A}) = \operatorname{tr}(\boldsymbol{A}')$.
2. $\operatorname{tr}(c\boldsymbol{A}) = c\operatorname{tr}(\boldsymbol{A})$, for c scalar.
3. $\operatorname{tr}(\boldsymbol{A}+\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B})$, for any conformable $\boldsymbol{B}$.
4. For any $\boldsymbol{B} \in \mathbb{R}^{n \times m}$ and $\boldsymbol{C} \in \mathbb{R}^{m \times n}$, $\operatorname{tr}(BC) = \operatorname{tr}(CB)$.
5. $\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i$, where λ_i, $i = 1, \ldots, n$, are the eigenvalues of $\boldsymbol{A}$.

9 Two Useful Matrix Decompositions

Eigendecomposition. Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with n linearly independent eigenvectors. There exists a full rank matrix $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ such that

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1},$$

where $\boldsymbol{P}$ is such that its i-th column is the i-th eigenvector of $\boldsymbol{A}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues of $\boldsymbol{A}$; that is, $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$.

When $\boldsymbol{A}$ is symmetric, in particular, since eigenvectors of real symmetric matrices are orthogonal, by normalizing eigenvectors to make them orthonormal one can always construct an orthogonal $\boldsymbol{P}$, so that $\boldsymbol{P}^{-1} = \boldsymbol{P}'$. This follows from the fact that a square matrix with orthonormal columns is always orthogonal.[5] Therefore, for symmetric $\boldsymbol{A}$ one can always find $\boldsymbol{P}$ such that $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$, with $\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{I}_n$.

Matrix square root. Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a positive definite matrix ($\boldsymbol{A} \succ 0$). There exists $\boldsymbol{B}$ such that

$$\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}'.$$

We refer to $\boldsymbol{B}$ as the *matrix square root* of $\boldsymbol{A}$, denoted by $\boldsymbol{A}^{1/2}$. Please note that this is simply a notation; it does *not* imply that all entries of $\boldsymbol{A}$ are being square-rooted or anything of that sort. Matrix square roots are common in Econometrics. Variance matrices of vector random variables are typically positive definite and symmetric; it is standard practice to take the matrix square root of these variance matrices. Matrix square rooting is used, for example, in proving Aitken's theorem, which asserts that the generalized least squares estimator (GLS) is the best linear unbiased estimator (BLUE) under heteroskedastic errors.

[5] By *orthonormal columns* here I mean that every column has magnitude 1 (that is, for every column $\boldsymbol{v}_i$, $i = 1, \ldots, n$, $\|\boldsymbol{v}_i\| = 1$, where $\|\cdot\|$ denotes the usual L^2-norm), and all columns are mutually orthogonal (that is, for every $j \neq i$, $\boldsymbol{v}_i'\boldsymbol{v}_j = 0$).

10 The Multivariate Normal Distribution

The probability density function of a scalar random variable x following a *univariate* normal distribution $N(\mu, \sigma^2)$ is

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

What about the *multivariate* case — when $\boldsymbol{x}$ is a *vector* random variable?

Definition 14 (Multivariate standard normal distribution)**.** A k-dimensional random vector $\boldsymbol{x}$ is said to follow a *multivariate standard normal distribution* if it has density

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{\boldsymbol{x}'\boldsymbol{x}}{2}\right).$$

We denote $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I}_k)$.

Observe that $\boldsymbol{x}'\boldsymbol{x} = \sum_{i=1}^k x_i^2$, whence

$$\exp\left(-\frac{\boldsymbol{x}'\boldsymbol{x}}{2}\right) = \exp\left(-\frac{\sum_{i=1}^k x_i^2}{2}\right) = \prod_{i=1}^k \exp\left(-\frac{x_i^2}{2}\right).$$

Moreover,

$$\frac{1}{(2\pi)^{k/2}} = \prod_{i=1}^k \frac{1}{(2\pi)^{1/2}}.$$

Therefore,

$$f(\boldsymbol{x}) = \prod_{i=1}^k \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{x_i^2}{2}\right).$$

Observe that each term in the product above is an univariate standard normal density. This implies the following theorem.

Theorem 3. *If $\boldsymbol{x} = (x_1, x_2, \ldots, x_k) \sim N(\boldsymbol{0}, \boldsymbol{I}_k)$, then all the entries of $\boldsymbol{x}$ are independent and identically distributed by univariate standard normal distributions; that is, $x_i \sim N(0,1)$ for all $i = 1, \ldots, k$.*

This result has the practical implication that generating a draw from a k-dimensional random vector $\boldsymbol{x}$ is statistically equivalent to generating k draws from a scalar random variable $x \sim N(0,1)$. Unfortunately, this property holds only for standard normal distributions and is not preserved for general (nonstandard) normal distributions, as we will see below.

Definition 15 (Multivariate normal distribution)**.** Let $\boldsymbol{z} \sim N(0, \boldsymbol{I}_k)$ and $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{z}$, with $\boldsymbol{\mu} \in \mathbb{R}^q$ and $\boldsymbol{B} \in \mathbb{R}^{q \times k}$. The k-dimensional random vector $\boldsymbol{x}$ is said to follow a *multivariate normal distribution.* We denote $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \equiv \boldsymbol{B}\boldsymbol{B}'$. The density of $\boldsymbol{x}$ is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right).$$

Observe that $\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{z}] = \boldsymbol{\mu} + \boldsymbol{B}\mathbb{E}[\boldsymbol{z}] = \boldsymbol{\mu}$. Further, recall that the variance of a vector random variable $\boldsymbol{x}$ is defined by the operation $\text{var}(\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])']$ and observe that

$$\text{var}(\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})'] = \mathbb{E}[(\boldsymbol{B}\boldsymbol{z})(\boldsymbol{B}\boldsymbol{z})'] = \mathbb{E}[\boldsymbol{B}\boldsymbol{z}\boldsymbol{z}'\boldsymbol{B}] = \boldsymbol{B}\mathbb{E}[\boldsymbol{z}\boldsymbol{z}']\boldsymbol{B} = \boldsymbol{B}\boldsymbol{B}' = \boldsymbol{\Sigma}.$$

For this reason, $\boldsymbol{\mu}$ is called the mean of the random vector $\boldsymbol{x}$ and $\boldsymbol{\Sigma}$ the variance (or variance-covariance) matrix of $\boldsymbol{x}$. $\boldsymbol{\Sigma}$ is always square, symmetric and positive semi-definite. One can further verify that

$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) & \text{cov}(x_1, x_3) & \cdots & \text{cov}(x_1, x_k) \\ \text{cov}(x_1, x_2) & \text{var}(x_2) & \text{cov}(x_2, x_3) & \cdots & \text{cov}(x_2, x_k) \\ \text{cov}(x_1, x_3) & \text{cov}(x_2, x_3) & \text{var}(x_3) & \cdots & \text{cov}(x_3, x_k) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{cov}(x_1, x_k) & \text{cov}(x_2, x_k) & \text{cov}(x_3, x_k) & \cdots & \text{var}(x_k) \end{bmatrix}.$$

It is possible to show that if $\boldsymbol{x}$ follows a multivariate normal distribution, then each element of $\boldsymbol{x}$ follows a univariate normal distribution. However, unlike in the case of the standard normal distribution, these univariate normal distributions will not necessarily be independent and/or identically distributed.

Indeed, if $\boldsymbol{\Sigma}$ is diagonal, then x_i, $i = 1, \ldots, k$, are independent, but not necessarily identically distributed. However, if $\mu_i = \mu_j =: \mu_0$ for all i, j, $\boldsymbol{\Sigma}$ is diagonal, and $\sigma_i^2 = \sigma_j^2 =: \sigma_0^2$, then x_i, $i = 1, \ldots, k$, are independent and identically distributed as $x_i \sim N(\mu_0, \sigma_0^2)$. I shall demonstrate the former result; the latter is left to the reader.

Observe that diagonality of $\boldsymbol{\Sigma}$ implies

$$\det(\boldsymbol{\Sigma}) = \prod_{i=1}^{k} \sigma_i^2, \quad \text{and} \quad (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = \sum_{i=1}^{k} \frac{(x_i - \mu_i)^2}{\sigma_i^2}.$$

Thus

$$f(\boldsymbol{x}) = \prod_{i=1}^{k} \frac{1}{(2\pi\sigma_i^2)^{1/2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right).$$

Notice that each term in the product above corresponds to a univariate normal density for a $N(\mu_i, \sigma_i^2)$ distribution. This implies that the elements of $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are independent, but not necessarily identically distributed. We have thus established the following theorem.

Theorem 4. *If $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ is a k-dimensional normally distributed vector random variable, then x_i, $i = 1, \ldots, k$, are independent if and only if cov$(x_i, x_j) = 0$ for all $i \neq j$, $j = 1, \ldots, k$.*

Recall that independence between random variables always implies zero correlation. However, in general, zero correlation between random variables does not always imply independence. Theorem 4 states, however, that when all random variables under consideration are normally distributed, the converse is also true.

Below, I present two additional useful results related to multivariate normal distributions.

Theorem 5. *If $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{B}x$, then $\boldsymbol{y} \boldsymbol{\sim} \boldsymbol{N}(\boldsymbol{a} + \boldsymbol{B\mu}, \boldsymbol{B\Sigma B'})$.*

Theorem 6. *If $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ is a k-dimensional random vector, then $\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \sim \boldsymbol{\chi}^2_k$.*

11 A Quick Glimpse into Matrix Differential Calculus

11.1 Scalar-by-vector derivative

Let $\boldsymbol{x} \in \mathbb{R}^k$ and $f : \mathbb{R}^k \to \mathbb{R}$ be a scalar function. Define the *scalar-by-vector derivative* as

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \equiv \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \frac{\partial f(\boldsymbol{x})}{\partial x_2}, \cdots, \frac{\partial f(\boldsymbol{x})}{\partial x_k} \right)'.$$

Consider the following two very simple scalar-by-vector derivative rules for quadratic forms and dot products:

$$\frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{x}, \quad \text{and} \quad \frac{\partial \boldsymbol{x}'\boldsymbol{v}}{\partial \boldsymbol{x}} = \boldsymbol{v}. \tag{8}$$

The first rule follows from the observation that $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{j=1}^k \sum_{j=1}^k a_{ij} x_i x_j$, whence

$$\frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \begin{bmatrix} \sum_{j=1}^k a_{1j} x_j + \sum_{i=1}^k a_{i1} x_i \\ \vdots \\ \sum_{j=1}^k a_{kj} x_j + \sum_{i=1}^k a_{ik} x_i \end{bmatrix} = (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{x}.$$

The second rule follows from the observation that $\boldsymbol{x}'\boldsymbol{v} = \sum_{i=1}^k x_i v_i$, whence

$$\frac{\partial \boldsymbol{x}'\boldsymbol{v}}{\partial \boldsymbol{x}} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix} = \boldsymbol{v}.$$

You may recognize that these scalar-by-vector operations are simply standard gradients from multivariate calculus. The point here is that the gradient of vector/matrix forms can be neatly represented with respect to their vector/matrix factors. Naturally, you don't want to recompute all these summations each time you take derivatives of quadratic forms or dot products. Moreover, during matrix-algebraic calculations, it's preferable to maintain consistency in matrix/vector representation rather than mixing summation and matrix representations. For these reasons, it's beneficial to become accustomed to the direct vector/matrix representations of these gradients, such as the final representations in (8).

The next example illustrates how even the two very simple rules discussed above can be highly useful in practice.

Example 2. Consider again the linear model as in (2), but now in its full matrix representation (3), which I repeat here for readability.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{9}$$

We know that the OLS estimator minimizes the sum of squared errors. That is,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} \boldsymbol{u}'\boldsymbol{u} \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}. \qquad (10)
\end{aligned}$$

The final equality follows from the fact that $\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}$ are scalars, whence equal to their own transposes. Now, to find the argmin, observe that the objective function is convex, so $\hat{\boldsymbol{\beta}}$ is the solution to the first-order conditions associated with (10). Using the rules (8), we can express these first-order conditions as

$$\begin{aligned}
\frac{\partial \boldsymbol{u}'\boldsymbol{u}}{\partial \boldsymbol{\beta}} &= -2\boldsymbol{X}'\boldsymbol{Y} + (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} \qquad (11) \\
&= -2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}. \qquad (12)
\end{aligned}$$

Then, $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$ and, provided $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular,

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

Using matrix calculus, we have derived the OLS estimator without any need for summations or standard scalar algebra. △

11.2 Vector-by-vector derivative

Let $\boldsymbol{x} \in \mathbb{R}^k$ and $\boldsymbol{f} : \mathbb{R}^k \to \mathbb{R}^n$ be a vector function. Define the *vector-by-vector* derivative as

$$\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}'} \equiv \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_k} \\ \frac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_2(\boldsymbol{x})}{\partial x_k} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_n(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_n(\boldsymbol{x})}{\partial x_k} \end{bmatrix}.$$

Observe that for the special case of a scalar function (that is, when $n = 1$) we have

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}'} = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{f_1(\boldsymbol{x})}{\partial x_k} \end{bmatrix} = \left(\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \right)'.$$

The notation $\partial \boldsymbol{f}(\boldsymbol{x})/\partial \boldsymbol{x}'$ has the advantage of being consistent with the dimensions of its resulting matrix: we differentiate n elements of a column vector, $\boldsymbol{f}(\boldsymbol{x})$, with respect to k elements of a row vector, $\boldsymbol{x}'$, and this gives us a $n \times k$ matrix. It is important to notice, however, that this is *just* a handy notation; it is not conceptual.

11.3 Second-order scalar-by-vector derivative

The first-order derivative of a scalar function is a vector. A direct consequence of the definition of vector-by-vector derivative is that the second-order derivative of a scalar function is a matrix. We define the second-order scalar-by-vector derivative as

$$\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}'} \equiv \frac{\partial}{\partial \boldsymbol{x}} \left(\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \right)' = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_k} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_k} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_k \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_k \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_k \partial x_k} \end{bmatrix}$$

You may recognize that these scalar-by-vector operations are simply standard hessians from multivariate calculus.

It's important to note that while scalar-by-vector and vector-by-vector derivatives may seem straightforward as they resemble gradients and hessians, matrix calculus can quickly become complex when considering other types of derivatives, such as scalar-by-matrix and matrix-by-matrix derivatives. All these derivatives can be well-defined and possess useful vector/matrix-algebraic properties. I won't get into the details of these other types of derivatives here. For a summary, you may find this Wikipedia entry to be a good initial resource. For a comprehensive reference on matrix differential calculus, I highly recommend Magnus and Neudecker [2019].[6]

11.4 Multivariate Taylor Expansions

Taylor expansions are powerful tools in statistics, used to derive asymptotic properties of estimators resulting from nonlinear estimating equations, proving the Delta Method and determining the asymptotic distribution of test statistics.

Let $\boldsymbol{x} \in \mathbb{R}^k$ and $f : \mathbb{R}^k \to \mathbb{R}$ be a scalar function. The first-order scalar-by-vector Taylor expansion is given by

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\boldsymbol{x} - \boldsymbol{x}_0)' \frac{\partial f(\boldsymbol{x}_0)}{\partial \boldsymbol{x}} + o\left(\|\boldsymbol{x} - \boldsymbol{x}_0\|\right) \text{ as } \boldsymbol{x} \to \boldsymbol{x}_0.$$

[6] For a discussion on how matrix calculus can quickly become confusing, check out the discussion on "derisatives" in Chapter 9.3 of the 2019 third edition of the book. It's funny.

The second-order scalar-by-vector Taylor expansion is given by

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\boldsymbol{x} - \boldsymbol{x}_0)' \frac{\partial f(\boldsymbol{x}_0)}{\partial \boldsymbol{x}} + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)' \frac{\partial f(\boldsymbol{x}_0)}{\partial \boldsymbol{x} \partial \boldsymbol{x}'}(\boldsymbol{x} - \boldsymbol{x}_0) + o\left(\|\boldsymbol{x} - \boldsymbol{x}_0\|^2\right) \text{ as } \boldsymbol{x} \to \boldsymbol{x}_0.$$

Let $\boldsymbol{x} \in \mathbb{R}^k$ and $\boldsymbol{f} : \mathbb{R}^k \to \mathbb{R}^n$ be a vector function. The second-order vector-by-vector Taylor expansion is given by

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0) + \frac{\partial \boldsymbol{f}(\boldsymbol{x}_0)}{\partial \boldsymbol{x}'}(\boldsymbol{x} - \boldsymbol{x}_0) + o(\|\boldsymbol{x} - \boldsymbol{x}_0\|) \text{ as } \boldsymbol{x} \to \boldsymbol{x}_0.$$

Here, $o(\cdot)$ is the little-o notation, defined as follows.

Definition 16 (Little-o notation for limits of functions)**.** Assume $g(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \neq \boldsymbol{x}_0$ in some neighborhood containing $\boldsymbol{x}_0$. The notation

$$f(\boldsymbol{x}) = o(g(\boldsymbol{x})) \text{ as } \boldsymbol{x} \to \boldsymbol{x}_0 \quad \text{means that} \quad \lim_{\boldsymbol{x} \to \boldsymbol{x}_0} \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} = 0.$$

The symbol $f(\boldsymbol{x}) = o(g(\boldsymbol{x}))$ is read "$f(\boldsymbol{x})$ is little-oh of $g(\boldsymbol{x})$", or "$f(\boldsymbol{x})$ is of smaller order than $g(\boldsymbol{x})$", and it is intended to convey the idea that for $\boldsymbol{x}$ near $\boldsymbol{x}_0$, $f(\boldsymbol{x})$ is small compared to $g(\boldsymbol{x})$; or, in other words, that $f(\boldsymbol{x})$ goes faster to zero than $g(\boldsymbol{x})$.

A similar definition is possible for sequences of numbers.

Definition 17 (Little-o notation for limits of sequences)**.** The notation

$$x_n = o(a_n) \quad \text{means that} \quad \lim_{n \to \infty} \frac{x_n}{a_n} = 0.$$

More importantly for our purposes, a similar definition is possible for sequences of random variables under the notion of convergence in probability.

Definition 18. Let Z_n be a sequence of random variables and a_n a sequence of constants. The notation

$$Z_n = o_p(a_n) \quad \text{means that} \quad \frac{Z_n}{a_n} \xrightarrow{p} 0.$$

Observe that, in particular, when $a_n = o(1)$ we have that $a_n \to 0$; similarly, when $Z_n = o_p(1)$ we have that $Z_n \xrightarrow{p} 0$.

In statistical applications, Taylor expansions are typically applied to sequences of random variables. Thus, the $o(\cdot)$ terms in the Taylor expansions presented above are replaced by $o_p(1)$ terms. Whether applying Taylor expansions to sequences of random variables is valid is a subtle discussion that I want to avoid here (see, e.g., Feng et al. [2013], Yang and Zhou [2021], Patriota [2019]). Another subtle point is whether, assuming Taylor expansions can indeed be applied to sequences of random variables, the Taylor expansion errors are in fact $o_p(1)$. Most textbooks, even advanced ones, typically ignore these issues and apply Taylor expansions to random variables without mentioning these discussions. It is common practice to ignore expansion errors and omit $o_p(1)$ terms.

12 Multivariate Asymptotics

Theorem 7 (Multivariate Weak Law of Large Numbers). *If $X_i \in \mathbb{R}^m$ are independent and identically distributed and $\mathbb{E}\|X\| < \infty$, then as $n \to \infty$,*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X].$$

Theorem 8 (Multivariate Strong Law of Large Numbers). *If $X_i \in \mathbb{R}^m$ are independent and identically distributed and $\mathbb{E}\|X\| < \infty$, then as $n \to \infty$,*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}[X].$$

The Strong Law of Large Numbers is more elegant than the Weak Law of Large Numbers; however, for most practical purposes the Weak Law is sufficient. Thus in econometrics we primarily use the Weak Law.

Theorem 9 (Multivariate Central Limit Theorem). *If $X_i \in \mathbb{R}^m$ are i.i.d. and $\mathbb{E}\|X\|^2 < \infty$, then as $n \to \infty$*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma),$$

where $\mu \equiv \mathbb{E}[X]$ and $\Sigma \equiv \mathbb{E}[(X-\mu)(X-\mu)']$.

Theorem 10 (Multivariate Continuous Mapping Theorem). *If $Z_n \to_d Z$ as $n \to \infty$ and $h : \mathbb{R}^m \to \mathbb{R}^k$ has the set of discontinuity points D_h such that $P[Z \in D_h] = 0$, then $h(Z_n) \xrightarrow{d} h(Z)$ as $n \to \infty$.*

A special case of the Continuous Mapping Theorem is known as Slutsky's Theorem.

Theorem 11 (Slutsky's Theorem). *If $Z_n \xrightarrow{d} Z$ and $c_n \xrightarrow{p} c$ as $n \to \infty$, then*

1. $Z_n + c_n \xrightarrow{d} Z + c$
2. $Z_n c_n \xrightarrow{d} Zc$
3. $\frac{Z_n}{c_n} \xrightarrow{d} \frac{Z}{c}$,

provided $c \neq 0$.

Notice that both the Continuous Mapping Theorem (CMT) and Slutky's Theorem hold, in particular, for convergence in probability.

Theorem 12 (Multivariate Delta Method). *Let $\theta \in \mathbb{R}^k$. If $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \xi$ and $h : \mathbb{R}^k \to \mathbb{R}$ is continuously differentiable in a neighborhood of θ then as $n \to \infty$*

$$\sqrt{n}(h(\hat{\theta}) - h(\theta)) \xrightarrow{d} \frac{\partial h(\theta)}{\partial \theta'}\xi.$$

In particular, if $\xi \sim N(0, \Sigma)$ we have $\xrightarrow{d} N\left(0, \frac{\partial h(\theta)}{\partial \theta'}\xi\frac{\partial h(\theta)}{\partial \theta}\right)$.

References

Changyong Feng, Hongyue Wang, Yu Han, Yinglin Xia, and Xin M Tu. The mean value theorem and taylor's expansion in statistics. *The American Statistician*, 67(4):245–248, 2013.

J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Series in Probability and Statistics. Wiley, 2019. ISBN 9781119541196. URL https://books.google.com.br/books?id=9jmNDwAAQBAJ.

Alexandre Galvão Patriota. On the mean value theorem for estimating functions. *The American Statistician*, 2019.

Yifan Yang and Xiaoyu Zhou. A note on taylor's expansion and mean value theorem with respect to a random variable. *arXiv preprint arXiv:2102.10429*, 2021.