Linear Models

Claudio Agostinelli
claudio@unive.it
DAIS
Ca' Foscari University
San Giobbe, Cannaregio 873, Venezia
Tel. 041 2347446, Fax. 041 2347444
http://www.dst.unive.it/~claudio

February 23, 2012

Contents

-3 Matrix Algebra

In this part we introduce essential material from Matrix Algebra. Most of the material is taken from Beezer [2006] which I strongly suggest you to download and study. Some arguments are missed in this book, you could find the remain topics in Schott [1997] and in the Appendix A of Rao and Toutenburg [1995]. Most of the theorems are followed by their proof. Please look into it. Also, if you feel voluntary to add the missed proofs please contact me.

-3.1 Matrices

Definition 1 (Matrix) An $m \times n$ matrix A is a rectangular array of elements in m rows and n columns. We indicate an $m \times n$ matrix by writing $A : m \times n$ or $A_{m \times n}$. Let A_{ij} be the element in the ith row and the jth column of A. Then A may be represented as

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & & & & \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}$$

and $a_{ij} = A_{ij}$. △

Example 2

$$B = \begin{bmatrix} -1 & 2 & 5 & 3 \\ 1 & 0 & -6 & 1 \\ -4 & 2 & 2 & -2 \end{bmatrix}$$

is a matrix with $m = 3$ rows and $n = 4$ columns. We can say that $B_{2,3} = -6$ while $B_{3,4} = -2$. ⊠

- A matrix with $m = n$ rows and columns is called **square matrix** ;
- A square matrix having zeros as elements below (above) the diagonale is called **upper (lower) triangular matrix**;
- To emphasize the situation when a matrix is not square, we will call it **rectangular** .

Definition 3 (Matrix Equality) The $m \times n$ matrices A and B are **equal**, written $A = B$ provided $A_{ij} = B_{ij}$ for all $1 \leq i \leq m$, $1 \leq j \leq n$. △

Definition 4 (Matrix partition) A matrix A is said to be **partitioned** if its elements are arranged in submatrices. $\triangle$

Example 5 (Matrix partition)

$$A_{m\times n} = \begin{bmatrix} B_{m\times r} & C_{m\times s} \end{bmatrix} \text{ with } r+s=m$$

or

$$A_{m\times n} = \begin{bmatrix} B_{r\times(n-s)} & C_{r\times s} \\ D_{(m-r)\times(n-s)} & E_{(m-r)\times s} \end{bmatrix} \qquad \boxtimes$$

-3.2 Vectors

Definition 6 (Column Vector) An $m \times 1$ matrix $\boldsymbol{a}$ is said to be an m-vector and written as a column

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_m \end{bmatrix} \qquad \triangle$$

Definition 7 (Column Vector Equality) The vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are **equal**, written $\boldsymbol{u} = \boldsymbol{v}$ provided that

$$\boldsymbol{u}_i = \boldsymbol{v}_i \qquad\qquad 1 \le i \le m \qquad \triangle$$

-3.3 Operations with vectors

Definition 8 (Column Vector Addition) Given the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ the **sum** of $\boldsymbol{u}$ and $\boldsymbol{v}$ is the vector $\boldsymbol{u}+\boldsymbol{v}$ defined by

$$\boldsymbol{u}+\boldsymbol{v}_i = \boldsymbol{u}_i + \boldsymbol{v}_i \qquad\qquad 1 \le i \le m \qquad \triangle$$

So vector addition takes two vectors of the same size and combines them (in a natural way!) to create a new vector of the same size. Notice that this definition is required, even if we agree that this is the obvious, right, natural or correct way to do it. Notice too that the symbol '+' is being recycled. We all know how to add *numbers*, but now we have the same symbol extended to double-duty and we use it to indicate how to add two new objects, vectors. And this definition of our new meaning is built on our previous meaning of addition via the expressions $u_i + v_i$. Think about your objects, especially when doing proofs.

Example 9 (Vector Addition) If

$$\boldsymbol{u} = \begin{bmatrix} 2 \\ -3 \\ 4 \\ 2 \end{bmatrix} \qquad \boldsymbol{v} = \begin{bmatrix} -1 \\ 5 \\ 2 \\ -7 \end{bmatrix}$$

then

$$\boldsymbol{u} + \boldsymbol{v} = \begin{bmatrix} 2 \\ -3 \\ 4 \\ 2 \end{bmatrix} + \begin{bmatrix} -1 \\ 5 \\ 2 \\ -7 \end{bmatrix} = \begin{bmatrix} 2 + (-1) \\ -3 + 5 \\ 4 + 2 \\ 2 + (-7) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 6 \\ -5 \end{bmatrix}. \qquad \boxtimes$$

Definition 10 (Column Vector Scalar Multiplication) Given the vector $\boldsymbol{u}$ and the scalar $\alpha \in \mathbb{R}$, the **scalar multiple** of $\boldsymbol{u}$ by α, $\alpha\boldsymbol{u}$ is defined by

$$\alpha\boldsymbol{u}_i = \alpha\boldsymbol{u}_i \qquad 1 \leq i \leq m \qquad \triangle$$

Notice that we are doing a kind of multiplication here, but we are *defining* a new type, perhaps in what appears to be a natural way. We use juxtaposition (smashing two symbols together side-by-side) to denote this operation rather than using a symbol like we did with vector addition. So this can be another source of confusion. When two symbols are next to each other, are we doing regular old multiplication, the kind we've done for years, or are we doing scalar vector multiplication, the operation we just defined? Think about your objects — if the first object is a scalar, and the second is a vector, then it *must* be that we are doing our new operation, and the *result* of this operation will be another vector.

Example 11 (Vector scalar multiplication) If

$$\boldsymbol{u} = \begin{bmatrix} 3 \\ 1 \\ -2 \\ 4 \\ -1 \end{bmatrix}$$

and $\alpha = 6$, then

$$\alpha\boldsymbol{u} = 6\begin{bmatrix} 3 \\ 1 \\ -2 \\ 4 \\ -1 \end{bmatrix} = \begin{bmatrix} 6(3) \\ 6(1) \\ 6(-2) \\ 6(4) \\ 6(-1 \end{bmatrix} = \begin{bmatrix} 18 \\ 6 \\ -12 \\ 24 \\ -6 \end{bmatrix}. \qquad \boxtimes$$

Definition 12 (Vector Linear Combination) Given n vectors $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, $\boldsymbol{u}_3$, $\ldots$, $\boldsymbol{u}_n$ from $\mathbb{R}^m$ and n scalars α_1, α_2, α_3, $\ldots$, α_n, their **linear combination** is the vector

$$\alpha_1\boldsymbol{u}_1 + \alpha_2\boldsymbol{u}_2 + \alpha_3\boldsymbol{u}_3 + \cdots + \alpha_n\boldsymbol{u}_n. \qquad \triangle$$

Example 13 (Vector Linear Combination) Suppose that

$$\alpha_1 = 1 \qquad \alpha_2 = -4 \qquad \alpha_3 = 2 \qquad \alpha_4 = -1$$

and

$$\boldsymbol{u}_1 = \begin{bmatrix} 2 \\ 4 \\ -3 \\ 1 \\ 2 \\ 9 \end{bmatrix} \qquad \boldsymbol{u}_2 = \begin{bmatrix} 6 \\ 3 \\ 0 \\ -2 \\ 1 \\ 4 \end{bmatrix} \qquad \boldsymbol{u}_3 = \begin{bmatrix} -5 \\ 2 \\ 1 \\ 1 \\ -3 \\ 0 \end{bmatrix} \qquad \boldsymbol{u}_4 = \begin{bmatrix} 3 \\ 2 \\ -5 \\ 7 \\ 1 \\ 3 \end{bmatrix}$$

then their linear combination $\alpha_1\boldsymbol{u_1} + \alpha_2\boldsymbol{u_2} + \alpha_3\boldsymbol{u_3} + \alpha_4\boldsymbol{u_4}$ is

$$\begin{aligned} (1)\begin{bmatrix} 2 \\ 4 \\ -3 \\ 1 \\ 2 \\ 9 \end{bmatrix} + (-4)\begin{bmatrix} 6 \\ 3 \\ 0 \\ -2 \\ 1 \\ 4 \end{bmatrix} + (2)\begin{bmatrix} -5 \\ 2 \\ 1 \\ 1 \\ -3 \\ 0 \end{bmatrix} + (-1)\begin{bmatrix} 3 \\ 2 \\ -5 \\ 7 \\ 1 \\ 3 \end{bmatrix} \\ = \begin{bmatrix} 2 \\ 4 \\ -3 \\ 1 \\ 2 \\ 9 \end{bmatrix} + \begin{bmatrix} -24 \\ -12 \\ 0 \\ 8 \\ -4 \\ -16 \end{bmatrix} + \begin{bmatrix} -10 \\ 4 \\ 2 \\ 2 \\ -6 \\ 0 \end{bmatrix} + \begin{bmatrix} -3 \\ -2 \\ 5 \\ -7 \\ -1 \\ -3 \end{bmatrix} \\ = \begin{bmatrix} -35 \\ -6 \\ 4 \\ 4 \\ -9 \\ -10 \end{bmatrix}. \end{aligned}$$

A different linear combination, of the same set of vectors, can be formed with different scalars. Take

$$\beta_1 = 3 \qquad \beta_2 = 0 \qquad \beta_3 = 5 \qquad \beta_4 = -1$$

and form the linear combination

$$\beta_1\boldsymbol{u_1}+\beta_2\boldsymbol{u_2}+\beta_3\boldsymbol{u_3}+\beta_4\boldsymbol{u_4}=(3)\begin{bmatrix}2\\4\\-3\\1\\2\\9\end{bmatrix}+(0)\begin{bmatrix}6\\3\\0\\-2\\1\\4\end{bmatrix}+(5)\begin{bmatrix}-5\\2\\1\\1\\-3\\0\end{bmatrix}+(-1)\begin{bmatrix}3\\2\\-5\\7\\1\\3\end{bmatrix}$$

$$=\begin{bmatrix}6\\12\\-9\\3\\6\\27\end{bmatrix}+\begin{bmatrix}0\\0\\0\\0\\0\\0\end{bmatrix}+\begin{bmatrix}-25\\10\\5\\5\\-15\\0\end{bmatrix}+\begin{bmatrix}-3\\-2\\5\\-7\\-1\\-3\end{bmatrix}=\begin{bmatrix}-22\\20\\1\\1\\-10\\24\end{bmatrix}\boxtimes$$

Notice how we could keep our set of vectors fixed, and use different sets of scalars to construct different vectors. You might build a few new linear combinations of $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, $\boldsymbol{u}_3$, $\boldsymbol{u}_4$ right now. We'll be right here when you get back. What vectors were you able to create? Do you think you could create the vector

$$\boldsymbol{w}=\begin{bmatrix}13\\15\\5\\-17\\2\\25\end{bmatrix}$$

with a "suitable" choice of four scalars? Do you think you could create *any* possible vector from $\mathbb{R}^6$ by choosing the proper scalars? These last two questions are very fundamental, and time spent considering them *now* will prove beneficial later.

-3.4 System of Linear Equations

Example 14 (System of linear equations) As a vector equality, the following System of linear equations

$$\begin{aligned}x_1-x_2+2x_3&=1\\2x_1+x_2+x_3&=8\\x_1+x_2&=5\end{aligned}$$

can be written as

$$\begin{bmatrix} x_1 - x_2 + 2x_3 \\ 2x_1 + x_2 + x_3 \\ x_1 + x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \\ 5 \end{bmatrix}.$$

Now bust up the linear expressions on the left, first using vector addition,

$$\begin{bmatrix} x_1 \\ 2x_1 \\ x_1 \end{bmatrix} + \begin{bmatrix} -x_2 \\ x_2 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2x_3 \\ x_3 \\ 0x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \\ 5 \end{bmatrix}.$$

Rewrite each of these $n = 3$ vectors as a scalar multiple of a fixed vector, where the scalar is one of the unknown variables, converting the left-hand side into a linear combination

$$x_1 \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} + x_3 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \\ 5 \end{bmatrix}. \qquad \boxtimes$$

-3.5 Vector Space of Column Vectors

Definition 15 (Vector Space of Column Vectors) The vector space $\mathbb{R}^m$ is the set of all column vectors of size m with entries from the set of real numbers, $\mathbb{R}$. $\triangle$

Theorem 16 (Vector Space Properties of Column Vectors) Suppose that $\mathbb{R}^m$ is the set of column vectors of size m with addition and scalar multiplication. Then

- **ACC Additive Closure, Column Vectors** If $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$, then $\boldsymbol{u} + \boldsymbol{v} \in \mathbb{R}^m$.
- **SCC Scalar Closure, Column Vectors** If $\alpha \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^m$, then $\alpha\boldsymbol{u} \in \mathbb{R}^m$.
- **CC Commutativity, Column Vectors** If $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$, then $\boldsymbol{u}+\boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$.
- **AAC Additive Associativity, Column Vectors** If $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^m$, then $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$.
- **ZC Zero Vector, Column Vectors** There is a vector, $\mathbf{0}$, called the **zero vector**, such that $\boldsymbol{u} + \mathbf{0} = \boldsymbol{u}$ for all $\boldsymbol{u} \in \mathbb{R}^m$.

- **AIC Additive Inverses, Column Vectors** If $\boldsymbol{u} \in \mathbb{R}^m$, then there exists a vector $-\boldsymbol{u} \in \mathbb{R}^m$ so that $\boldsymbol{u} + (-\boldsymbol{u}) = \boldsymbol{0}$.
- **SMAC Scalar Multiplication Associativity, Column Vectors** If α, $\beta \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^m$, then $\alpha(\beta\boldsymbol{u}) = (\alpha\beta)\boldsymbol{u}$.
- **DVAC Distributivity across Vector Addition, Column Vectors** If $\alpha \in \mathbb{R}$ and $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{R}^m$, then $\alpha(\boldsymbol{u} + \boldsymbol{v}) = \alpha\boldsymbol{u} + \alpha\boldsymbol{v}$.
- **DSAC Distributivity across Scalar Addition, Column Vectors** If α, $\beta \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^m$, then $(\alpha + \beta)\boldsymbol{u} = \alpha\boldsymbol{u} + \beta\boldsymbol{u}$.
- **OC One, Column Vectors** If $\boldsymbol{u} \in \mathbb{R}^m$, then $1\boldsymbol{u} = \boldsymbol{u}$. □

Proof While some of these properties seem very obvious, they all require proof. However, the proofs are not very interesting, and border on tedious. We'll prove one version of distributivity very carefully, and you can test your proof-building skills on some of the others. We need to establish an equality, so we will do so by beginning with one side of the equality, apply various definitions and theorems (listed to the right of each step) to massage the expression from the left into the expression on the right.

Here we go with a proof of the Distributivity across Scalar Addition property. For $1 \leq i \leq m$,

$$\begin{aligned}
(\alpha + \beta)\boldsymbol{u}_i &= (\alpha + \beta)\boldsymbol{u}_i \\
&= \alpha\boldsymbol{u}_i + \beta\boldsymbol{u}_i && \text{Distributivity in } \mathbb{R} \\
&= \alpha\boldsymbol{u}_i + \beta\boldsymbol{u}_i \\
&= \alpha\boldsymbol{u} + \beta\boldsymbol{u}_i
\end{aligned}$$

Since the individual components of the vectors $(\alpha + \beta)\boldsymbol{u}$ and $\alpha\boldsymbol{u} + \beta\boldsymbol{u}$ are equal for *all* i, $1 \leq i \leq m$, the definition of equality between two vectors tells us the vectors are equal. ■

-3.6 Inner Products and Norms

Definition 17 (Inner Product) Given the vectors $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{R}^m$ the **inner product** of $\boldsymbol{u}$ and $\boldsymbol{v}$ is the scalar quantity in $\mathbb{R}$,

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}_1\boldsymbol{v}_1 + \boldsymbol{u}_2\boldsymbol{v}_2 + \boldsymbol{u}_3\boldsymbol{v}_3 + \cdots + \boldsymbol{u}_m\boldsymbol{v}_m = \sum_{i=1}^{m} \boldsymbol{u}_i\boldsymbol{v}_i \qquad \triangle$$

This operation is a bit different in that we begin with two vectors but produce a scalar. Computing one is straightforward.

Example 18 (Computing a inner product) The scalar product of

$$\boldsymbol{w} = \begin{bmatrix} 2 \\ 4 \\ -3 \\ 2 \\ 8 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{x} = \begin{bmatrix} 3 \\ 1 \\ 0 \\ -1 \\ -2 \end{bmatrix}$$

is

$$\langle \boldsymbol{w},\, \boldsymbol{x} \rangle = 2(3) + 4(1) + (-3)0 + 2(-1) + 8(-2) = -8. \qquad \boxtimes$$

In the case where the entries of our vectors are all real numbers the computation of the inner product may look familiar and be known to you as a **dot product** or **scalar product**. So you can view the inner product as a generalization of the scalar product to vectors from $\mathbb{R}^m$.

Theorem 19 (Inner Product and Vector Addition) Suppose $\boldsymbol{u}\,\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^m$. Then

1. $\langle \boldsymbol{u} + \boldsymbol{v},\, \boldsymbol{w} \rangle = \langle \boldsymbol{u},\, \boldsymbol{w} \rangle + \langle \boldsymbol{v},\, \boldsymbol{w} \rangle$
2. $\langle \boldsymbol{u},\, \boldsymbol{v} + \boldsymbol{w} \rangle = \langle \boldsymbol{u},\, \boldsymbol{v} \rangle + \langle \boldsymbol{u},\, \boldsymbol{w} \rangle$ $\qquad \square$

Proof The proofs of the two parts are very similar. We will prove part 2 and you can prove part 1.

$$\begin{aligned}
\langle \boldsymbol{u},\, \boldsymbol{v} + \boldsymbol{w} \rangle &= \sum_{i=1}^{m} \boldsymbol{u}_i \boldsymbol{v} + \boldsymbol{w}_i \\
&= \sum_{i=1}^{m} \boldsymbol{u}_i (\boldsymbol{v}_i + \boldsymbol{w}_i) \\
&= \sum_{i=1}^{m} \boldsymbol{u}_i \boldsymbol{v}_i + \boldsymbol{u}_i \boldsymbol{w}_i && \text{Distributivity in } \mathbb{R} \\
&= \sum_{i=1}^{m} \boldsymbol{u}_i \boldsymbol{v}_i + \sum_{i=1}^{m} \boldsymbol{u}_i \boldsymbol{w}_i && \text{Commutativity in } \mathbb{R} \\
&= \langle \boldsymbol{u},\, \boldsymbol{v} \rangle + \langle \boldsymbol{u},\, \boldsymbol{w} \rangle && \blacksquare
\end{aligned}$$

Theorem 20 (Inner Product and Scalar Multiplication) Suppose $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{R}^m$ and $\alpha \in \mathbb{R}$. Then

1. $\langle \alpha \boldsymbol{u},\, \boldsymbol{v} \rangle = \alpha \langle \boldsymbol{u},\, \boldsymbol{v} \rangle$
2. $\langle \boldsymbol{u},\, \alpha \boldsymbol{v} \rangle = \alpha \langle \boldsymbol{u},\, \boldsymbol{v} \rangle$ □

Proof The proofs of the two parts are very similar. We will prove part 2 and you can prove part 1.

$$\begin{aligned}
\langle \boldsymbol{u},\, \alpha \boldsymbol{v} \rangle &= \sum_{i=1}^{m} \boldsymbol{u}_i \alpha \boldsymbol{v}_i \\
&= \sum_{i=1}^{m} \boldsymbol{u}_i \alpha \boldsymbol{v}_i \\
&= \alpha \sum_{i=1}^{m} \boldsymbol{u}_i \boldsymbol{v}_i \qquad \text{Distributivity, Commutativity in } \mathbb{R} \\
&= \alpha \langle \boldsymbol{u},\, \boldsymbol{v} \rangle
\end{aligned}$$

■

Theorem 21 (Inner Product is (Anti-)Commutative) Suppose that $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors in $\mathbb{R}^m$. Then $\langle \boldsymbol{u},\, \boldsymbol{v} \rangle = \langle \boldsymbol{v},\, \boldsymbol{u} \rangle$. □

Proof

$$\begin{aligned}
\langle \boldsymbol{u},\, \boldsymbol{v} \rangle &= \sum_{i=1}^{m} \boldsymbol{u}_i \boldsymbol{v}_i \\
&= \sum_{i=1}^{m} \boldsymbol{u}_i\, \boldsymbol{v}_i \\
&= \sum_{i=1}^{m} \boldsymbol{v}_i \overline{\boldsymbol{u}_i} \qquad \text{Commutativity in } \mathbb{R} \\
&= \overline{\langle \boldsymbol{v},\, \boldsymbol{u} \rangle}
\end{aligned}$$

■

-3.7 Norms

Definition 22 (Norm of a Vector) The **norm** of the vector $\boldsymbol{u}$ is the scalar quantity in $\mathbb{R}$

$$\|\boldsymbol{u}\| = \sqrt{{\boldsymbol{u}_1}^2 + {\boldsymbol{u}_2}^2 + {\boldsymbol{u}_3}^2 + \cdots + {\boldsymbol{u}_m}^2} = \sqrt{\sum_{i=1}^{m} {\boldsymbol{u}_i}^2} \qquad \triangle$$

Computing a norm is also easy to do.

Example 23 (Computing the norm of some vectors) The norm of

$$\boldsymbol{u} = \begin{bmatrix} 3 \\ 1 \\ 2 \\ 2 \end{bmatrix}$$

is

$$\|\boldsymbol{u}\| = \sqrt{3^2 + 1^2 + 2^2 + 2^2} = \sqrt{9 + 1 + 4 + 4} = \sqrt{18} = 3\sqrt{2}.$$

The norm of

$$\boldsymbol{v} = \begin{bmatrix} 3 \\ -1 \\ 2 \\ 4 \\ -3 \end{bmatrix}$$

is

$$\|\boldsymbol{v}\| = \sqrt{3^2 + (-1)^2 + 2^2 + 4^2 + (-3)^2} = \sqrt{9 + 1 + 4 + 16 + 9} = \sqrt{39}. \quad \boxtimes$$

Notice how the norm of a vector is just the length of the vector. Inner products and norms are related by the following theorem.

Theorem 24 (Inner Products and Norms) Suppose that $\boldsymbol{u}$ is a vector in $\mathbb{R}^m$. Then $\|\boldsymbol{u}\|^2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle$. $\square$

Proof

$$\begin{aligned} \|\boldsymbol{u}\|^2 &= \left(\sqrt{\sum_{i=1}^{m} {\boldsymbol{u}_i}^2} \right)^2 \\ &= \sum_{i=1}^{m} {\boldsymbol{u}_i}^2 \\ &= \sum_{i=1}^{m} \boldsymbol{u}_i \boldsymbol{u}_i \\ &= \langle \boldsymbol{u}, \boldsymbol{u} \rangle \end{aligned}$$

■

This theorem says that the dot product of a vector with itself is equal to the length of the vector squared.

Theorem 25 (Positive Inner Products) Suppose that $\boldsymbol{u}$ is a vector in $\mathbb{R}^m$. Then $\langle \boldsymbol{u}, \boldsymbol{u} \rangle \geq 0$ with equality if and only if $\boldsymbol{u} = \boldsymbol{0}$. □

Proof

$$\langle \boldsymbol{u}, \boldsymbol{u} \rangle = {\boldsymbol{u}_1}^2 + {\boldsymbol{u}_2}^2 + {\boldsymbol{u}_3}^2 + \cdots + {\boldsymbol{u}_m}^2$$

Since each term is squared, every term is non–negative, and the sum must also be non–negative. The phrase, "with equality if and only if" means that we want to show that the statement $\langle \boldsymbol{u}, \boldsymbol{u} \rangle = 0$ (i.e. with equality) is equivalent ("if and only if") to the statement $\boldsymbol{u} = \boldsymbol{0}$.

If $\boldsymbol{u} = \boldsymbol{0}$, then it is a straightforward computation to see that $\langle \boldsymbol{u}, \boldsymbol{u} \rangle = 0$. In the other direction, assume that $\langle \boldsymbol{u}, \boldsymbol{u} \rangle = 0$. As before, $\langle \boldsymbol{u}, \boldsymbol{u} \rangle$ is a sum of non–negative terms. So we have

$$0 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle = {\boldsymbol{u}_1}^2 + {\boldsymbol{u}_2}^2 + {\boldsymbol{u}_3}^2 + \cdots + {\boldsymbol{u}_m}^2$$

Now we have a sum of squares equaling zero, so each term must be zero. Then by similar logic $\boldsymbol{u}_i = 0$. Thus every entry of $\boldsymbol{u}$ is zero and so $\boldsymbol{u} = \boldsymbol{0}$, as desired. ■

The conditions of the theorem are summarized by saying "the inner product is **positive definite**."

-3.8 Linearly Independent Sets of Vectors

Definition 26 (Relation of Linear Dependence for Column Vectors) Given a set of vectors $S = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \ldots, \boldsymbol{u}_n\}$, a true statement of the form

$$\alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2 + \alpha_3 \boldsymbol{u}_3 + \cdots + \alpha_n \boldsymbol{u}_n = \boldsymbol{0}$$

is a **relation of linear dependence** on S. If this statement is formed in a trivial fashion, i.e. $\alpha_i = 0$, $1 \leq i \leq n$, then we say it is the **trivial relation of linear dependence** on S. △

Definition 27 (Linear Independence of Column Vectors) The set of vectors $S = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \ldots, \boldsymbol{u}_n\}$ is **linearly dependent** if there is a relation of linear dependence on S that is not trivial. In the case where the *only* relation of linear dependence on S is the trivial one, then S is a **linearly independent** set of vectors. △

Notice that a relation of linear dependence is an *equation*. Though most of it is a linear combination, it is not a linear combination (that would be a

vector). Linear independence is a property of a *set* of vectors. It is easy to take a set of vectors, and an equal number of scalars, *all zero*, and form a linear combination that equals the zero vector. When the easy way is the *only* way, then we say the set is linearly independent. Here's a couple of examples.

Definition 28 (Span of a Set of Column Vectors) Given a set of vectors $S = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \ldots, \boldsymbol{u}_p\}$, their **span**, $\langle S \rangle$, is the set of all possible linear combinations of $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \ldots, \boldsymbol{u}_p$. Symbolically,

$$\begin{aligned}\langle S \rangle &= \{\alpha_1\boldsymbol{u}_1 + \alpha_2\boldsymbol{u}_2 + \alpha_3\boldsymbol{u}_3 + \cdots + \alpha_p\boldsymbol{u}_p \mid \alpha_i \in \mathbb{C},\ 1 \leq i \leq p\} \\ &= \left\{ \sum_{i=1}^{p} \alpha_i\boldsymbol{u}_i \;\middle|\; \alpha_i \in \mathbb{C},\ 1 \leq i \leq p \right\}\end{aligned}$$ △

The span is just a set of vectors, though in all but one situation it is an infinite set. (Just when is it not infinite?) So we start with a finite collection of vectors S (t of them to be precise), and use this finite set to describe an infinite set of vectors, $\langle S \rangle$. Confusing the *finite* set S with the *infinite* set $\langle S \rangle$ is one of the most pervasive problems in understanding introductory linear algebra. We will see this construction repeatedly, so let's work through some examples to get comfortable with it. The most obvious question about a set is if a particular item of the correct type is in the set, or not.

Definition 29 (Basis) Suppose V is a vector space. Then a subset $S \subseteq V$ is a **basis** of V if it is linearly independent and spans V. △

So, a basis is a linearly independent spanning set for a vector space. The requirement that the set spans V insures that S has enough raw material to build V, while the linear independence requirement insures that we do not have any more raw material than we need.

Definition 30 (Null Space of a Matrix) The **null space** of a matrix A, denoted $\mathcal{N}(A)$, is the set of all the vectors that are solutions to the homogeneous system $\mathcal{LS}(A, \mathbf{0})$, that is,

$$\mathcal{N}(A) = \{\boldsymbol{x} \in \mathbb{R}^n \text{ and } A\boldsymbol{x} = \mathbf{0}\} \subset \mathbb{R}^n$$ △

Definition 31 (Range Space of a Matrix) The **range space** of a matrix A, denoted $\mathcal{R}(A)$, is the vector space spanned by the column vectors of A, that is,

$$\mathcal{R}(A) = \{\boldsymbol{z} : \boldsymbol{z} = A\boldsymbol{x}; \boldsymbol{x} \in \mathbb{R}^n\} \subset \mathbb{R}^m$$ △

-3.9 Orthogonal Vectors

"Orthogonal" is a generalization of "perpendicular." You may have used mutually perpendicular vectors in a physics class, or you may recall from a calculus class that perpendicular vectors have a zero dot product. We will now extend these ideas into the realm of higher dimensions.

Definition 32 (Orthogonal Vectors) A pair of vectors, $\boldsymbol{u}$ and $\boldsymbol{v}$, from $\mathbb{R}^m$ are **orthogonal** if their inner product is zero, that is, $\langle \boldsymbol{u},\, \boldsymbol{v} \rangle = 0$. $\triangle$

Example 33 (Two orthogonal vectors) The vectors

$$\boldsymbol{u} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad\qquad \boldsymbol{v} = \begin{bmatrix} -\frac{3}{2} \\ 2 \\ 0 \\ \frac{3}{2} \end{bmatrix}$$

are orthogonal since

$$\begin{aligned} \langle \boldsymbol{u},\, \boldsymbol{v} \rangle &= 1(-\frac{3}{2}) + 0(2) + 0(0) + 1(\frac{3}{2}) \\ &= 0. \end{aligned}$$

$\boxtimes$

We extend this definition to whole sets by requiring vectors to be pairwise orthogonal. Despite using the same word, careful thought about what objects you are using will eliminate any source of confusion.

Definition 34 (Orthogonal Set of Vectors) Suppose that $S = \{\boldsymbol{u}_1,\, \boldsymbol{u}_2,\, \boldsymbol{u}_3,\, \ldots,\, \boldsymbol{u}_n\}$ is a set of vectors from $\mathbb{R}^m$. Then the set S is **orthogonal** if every pair of different vectors from S is orthogonal, that is $\langle \boldsymbol{u}_i,\, \boldsymbol{u}_j \rangle = 0$ whenever $i \neq j$.$\triangle$

Example 35 (Standard Unit Vectors are an Orthogonal Set) The standard unit vectors are the columns of the identity matrix (I). Computing the inner product of two distinct vectors, $\boldsymbol{e}_i$, $\boldsymbol{e}_j$, $i \neq j$, gives,

$$\begin{aligned} \langle \boldsymbol{e}_i,\, \boldsymbol{e}_j \rangle &= 0(0) + 0(0) + \cdots + 1(0) + \cdots + 0(1) + \cdots + 0(0) + 0(0) \\ &= 0 \end{aligned}$$

$\boxtimes$

Example 36 (An orthogonal set) The set

$$\{\boldsymbol{x}_1,\, \boldsymbol{x}_2,\, \boldsymbol{x}_3,\, \boldsymbol{x}_4\} = \left\{ \begin{bmatrix} 2 \\ 0 \\ 2 \\ 0 \end{bmatrix},\ \begin{bmatrix} \frac{3}{2} \\ 3 \\ -\frac{3}{2} \\ 0 \end{bmatrix},\ \begin{bmatrix} -\frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ 0 \end{bmatrix},\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{2} \end{bmatrix} \right\}$$

is an orthogonal set. Since the inner product is anti-commutative we can test pairs of different vectors in any order. If the result is zero, then it will also be zero if the inner product is computed in the opposite order. This means there are six pairs of different vectors to use in an inner product computation. We'll do two and you can practice your inner products on the other four.

$$\begin{aligned}\langle \boldsymbol{x}_1,\, \boldsymbol{x}_3\rangle &= 2(-\frac{1}{3}) + 0(\frac{1}{3}) + 2(\frac{1}{3}) + 0(0)\\ &= 0\end{aligned}$$

and

$$\begin{aligned}\langle \boldsymbol{x}_2,\, \boldsymbol{x}_4\rangle &= \frac{3}{2}(0) + 3(0) + (-\frac{3}{2})0 + 0(\frac{1}{2})\\ &= 0\end{aligned}$$

⊠

Theorem 37 (Orthogonal Sets are Linearly Independent) Suppose that $S = \{\boldsymbol{u}_1,\, \boldsymbol{u}_2,\, \boldsymbol{u}_3,\, \ldots,\, \boldsymbol{u}_n\}$ is an orthogonal set of nonzero vectors. Then S is linearly independent. □

Proof To prove linear independence of a set of vectors, we can appeal to its definition and begin with a relation of linear dependence,

$$\alpha_1\boldsymbol{u}_1 + \alpha_2\boldsymbol{u}_2 + \alpha_3\boldsymbol{u}_3 + \cdots + \alpha_n\boldsymbol{u}_n = \boldsymbol{0}.$$

Then, for every $1 \leq i \leq n$, we have

$$\begin{aligned}
0 &= 0\,\langle \boldsymbol{u}_i,\, \boldsymbol{u}_i\rangle\\
&= \langle 0\boldsymbol{u}_i,\, \boldsymbol{u}_i\rangle\\
&= \langle \boldsymbol{0},\, \boldsymbol{u}_i\rangle\\
&= \langle \alpha_1\boldsymbol{u}_1 + \alpha_2\boldsymbol{u}_2 + \alpha_3\boldsymbol{u}_3 + \cdots + \alpha_n\boldsymbol{u}_n,\, \boldsymbol{u}_i\rangle && \text{Relation of linear dependence}\\
&= \langle \alpha_1\boldsymbol{u}_1,\, \boldsymbol{u}_i\rangle + \langle \alpha_2\boldsymbol{u}_2,\, \boldsymbol{u}_i\rangle + \langle \alpha_3\boldsymbol{u}_3,\, \boldsymbol{u}_i\rangle + \cdots + \langle \alpha_n\boldsymbol{u}_n,\, \boldsymbol{u}_i\rangle\\
&= \alpha_1\,\langle \boldsymbol{u}_1,\, \boldsymbol{u}_i\rangle + \alpha_2\,\langle \boldsymbol{u}_2,\, \boldsymbol{u}_i\rangle + \alpha_3\,\langle \boldsymbol{u}_3,\, \boldsymbol{u}_i\rangle\\
&\quad + \cdots + \alpha_i\,\langle \boldsymbol{u}_i,\, \boldsymbol{u}_i\rangle + \cdots + \alpha_n\,\langle \boldsymbol{u}_n,\, \boldsymbol{u}_i\rangle\\
&= \alpha_1(0) + \alpha_2(0) + \alpha_3(0) + \cdots + \alpha_i\,\langle \boldsymbol{u}_i,\, \boldsymbol{u}_i\rangle + \cdots + \alpha_n(0) && \text{Orthogonal set}\\
&= \alpha_i\,\langle \boldsymbol{u}_i,\, \boldsymbol{u}_i\rangle
\end{aligned}$$

So we have $0 = \alpha_i\,\langle \boldsymbol{u}_i,\, \boldsymbol{u}_i\rangle$. However, since $\boldsymbol{u}_i \neq \boldsymbol{0}$ (the hypothesis said our vectors were nonzero), and hence we have $\langle \boldsymbol{u}_i,\, \boldsymbol{u}_i\rangle > 0$. So we must conclude that $\alpha_i = 0$ for all $1 \leq i \leq n$. But this says that S is a linearly independent set since the only way to form a relation of linear dependence is the trivial way, with all the scalars zero. Boom! ■

Definition 38 (OrthoNormal Set) Suppose $S = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \ldots, \boldsymbol{u}_n\}$ is an orthogonal set of vectors such that $\|\boldsymbol{u}_i\| = 1$ for all $1 \leq i \leq n$. Then S is an **orthonormal** set of vectors. △

Once you have an orthogonal set, it is easy to convert it to an orthonormal set — multiply each vector by the reciprocal of its norm, and the resulting vector will have norm 1. This scaling of each vector will not affect the orthogonality properties.

Example 39 (Orthonormal set, three vectors) The set

$$T = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3\} = \left\{ \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} \frac{3}{2} \\ 3 \\ -\frac{3}{2} \end{bmatrix}, \begin{bmatrix} -\frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix} \right\}$$

from the previous example is an orthogonal set. We compute the norm of each vector,

$$\|\boldsymbol{u}_1\| = 2\sqrt{2} \qquad \|\boldsymbol{u}_2\| = 3\sqrt{\frac{3}{2}} \qquad \|\boldsymbol{u}_3\| = \sqrt{\frac{1}{3}}$$

Converting each vector to a norm of 1, yields an orthonormal set,

$$\boldsymbol{w}_1 = \frac{\sqrt{2}}{4} \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix}$$

$$\boldsymbol{w}_2 = \frac{1}{3}\sqrt{\frac{2}{3}} \begin{bmatrix} \frac{3}{2} \\ 3 \\ -\frac{3}{2} \end{bmatrix}$$

$$\boldsymbol{w}_3 = \sqrt{3} \begin{bmatrix} -\frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}$$

⊠

-3.10 Matrix Operations

Definition 40 (Matrix Addition) Given the $m \times n$ matrices A and B, define the **sum** of A and B as an $m \times n$ matrix, written $A + B$, according to

$$A + B_{ij} = A_{ij} + B_{ij} \qquad 1 \leq i \leq m,\ 1 \leq j \leq n$$ △

So matrix addition takes two matrices of the same size and combines them (in a natural way!) to create a new matrix of the same size. Perhaps this is the "obvious" thing to do, but it doesn't relieve us from the obligation to state it carefully.

Example 41 (Addition of two matrices) If

$$A = \begin{bmatrix} 2 & -3 & 4 \\ 1 & 0 & -7 \end{bmatrix} \qquad B = \begin{bmatrix} 6 & 2 & -4 \\ 3 & 5 & 2 \end{bmatrix}$$

then

$$\begin{aligned} A + B &= \begin{bmatrix} 2 & -3 & 4 \\ 1 & 0 & -7 \end{bmatrix} + \begin{bmatrix} 6 & 2 & -4 \\ 3 & 5 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 2+6 & -3+2 & 4+(-4) \\ 1+3 & 0+5 & -7+2 \end{bmatrix} \\ &= \begin{bmatrix} 8 & -1 & 0 \\ 4 & 5 & -5 \end{bmatrix} \end{aligned}$$

⊠

Definition 42 (Matrix Scalar Multiplication) Given the $m \times n$ matrix A and the scalar $\alpha \in \mathbb{R}$, the **scalar multiple** of A is an $m \times n$ matrix, written αA and defined according to

$$\alpha A_{ij} = \alpha A_{ij} \qquad 1 \leq i \leq m,\ 1 \leq j \leq n$$

△

Example 43 (Scalar multiplication) If

$$A = \begin{bmatrix} 2 & 8 \\ -3 & 5 \\ 0 & 1 \end{bmatrix}$$

and $\alpha = 7$, then

$$\alpha A = 7 \begin{bmatrix} 2 & 8 \\ -3 & 5 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 7(2) & 7(8) \\ 7(-3) & 7(5) \\ 7(0) & 7(1) \end{bmatrix} = \begin{bmatrix} 14 & 56 \\ -21 & 35 \\ 0 & 7 \end{bmatrix}$$

⊠

Definition 44 (Zero Matrix) The $m \times n$ **zero matrix** is written as $\mathcal{O} = \mathcal{O}_{m \times n}$ and defined by $\mathcal{O}_{ij} = 0$, for all $1 \leq i \leq m$, $1 \leq j \leq n$. △

Theorem 45 (Vector Space Properties of Matrices) Suppose that M_{mn} is the set of all $m \times n$ matrices with matrix addition and scalar multiplication. Then

- **ACM Additive Closure, Matrices** If A, $B \in M_{mn}$, then $A + B \in M_{mn}$.

- **SCM Scalar Closure, Matrices** If $\alpha \in \mathbb{R}$ and $A \in M_{mn}$, then $\alpha A \in M_{mn}$.
- **CM Commutativity, Matrices** If A, $B \in M_{mn}$, then $A + B = B + A$.
- **AAM Additive Associativity, Matrices** If A, B, $C \in M_{mn}$, then $A + (B + C) = (A + B) + C$.
- **ZM Zero Vector, Matrices** There is a matrix, $\mathcal{O}$, called the **zero matrix**, such that $A + \mathcal{O} = A$ for all $A \in M_{mn}$.
- **AIM Additive Inverses, Matrices** If $A \in M_{mn}$, then there exists a matrix $-A \in M_{mn}$ so that $A + (-A) = \mathcal{O}$.
- **SMAM Scalar Multiplication Associativity, Matrices** If α, $\beta \in \mathbb{R}$ and $A \in M_{mn}$, then $\alpha(\beta A) = (\alpha\beta)A$.
- **DMAM Distributivity across Matrix Addition, Matrices** If $\alpha \in \mathbb{R}$ and A, $B \in M_{mn}$, then $\alpha(A + B) = \alpha A + \alpha B$.
- **DSAM Distributivity across Scalar Addition, Matrices** If α, $\beta \in \mathbb{R}$ and $A \in M_{mn}$, then $(\alpha + \beta)A = \alpha A + \beta A$.
- **OM One, Matrices** If $A \in M_{mn}$, then $1A = A$. □

Proof To prove $(\alpha + \beta)A = \alpha A + \beta A$, we need to establish the equality of two matrices. By definition we need to establish the equality of their entries, one-by-one. For *any* i and j, $1 \leq i \leq m$, $1 \leq j \leq n$,

$$\begin{aligned}
(\alpha + \beta)A_{ij} &= (\alpha + \beta)A_{ij} \\
&= \alpha A_{ij} + \beta A_{ij} && \text{Distributivity in } \mathbb{R} \\
&= \alpha A_{ij} + \beta A_{ij} \\
&= \alpha A + \beta A_{ij}
\end{aligned}$$

■

There are several things to notice here. (1) Each equals sign is an equality of numbers. (2) The two ends of the equation, being true for any i and j, allow us to conclude the equality of the matrices. (3) There are several plus signs, and several instances of juxtaposition. Identify each one, and state exactly what operation is being represented by each.

Definition 46 (Transpose of a Matrix) Given an $m \times n$ matrix A, its **transpose** is the $n \times m$ matrix A^t given by

$$A^t{}_{ij} = A_{ji}, \quad 1 \leq i \leq n, 1 \leq j \leq m. \qquad \triangle$$

Example 47 (Transpose of a 3 × 4 matrix) Suppose

$$D = \begin{bmatrix} 3 & 7 & 2 & -3 \\ -1 & 4 & 2 & 8 \\ 0 & 3 & -2 & 5 \end{bmatrix}.$$

We could formulate the transpose, entry-by-entry, using the definition. But it is easier to just systematically rewrite rows as columns (or vice-versa). The form of the definition given will be more useful in proofs. So we have

$$D^t = \begin{bmatrix} 3 & -1 & 0 \\ 7 & 4 & 3 \\ 2 & 2 & -2 \\ -3 & 8 & 5 \end{bmatrix} \qquad \boxtimes$$

Definition 48 (Symmetric Matrix) The matrix A is **symmetric** if $A = A^t$. $\triangle$

Example 49 (A symmetric 5 × 5 matrix) The matrix

$$E = \begin{bmatrix} 2 & 3 & -9 & 5 & 7 \\ 3 & 1 & 6 & -2 & -3 \\ -9 & 6 & 0 & -1 & 9 \\ 5 & -2 & -1 & 4 & -8 \\ 7 & -3 & 9 & -8 & -3 \end{bmatrix}$$

is symmetric. $\boxtimes$

Theorem 50 (Symmetric Matrices are Square) Suppose that A is a symmetric matrix. Then A is square. $\square$

Proof We start by specifying A's size, without assuming it is square, since we are trying to *prove* that, so we can't also assume it. Suppose A is an $m \times n$ matrix. Because A is symmetric, we know by definition that $A = A^t$. So, in particular, it requires that A and A^t must have the same size. The size of A^t is $n \times m$. Because A has m rows and A^t has n rows, we conclude that $m = n$, and hence A must be square. $\blacksquare$

Theorem 51 (Transpose and Matrix Addition) Suppose that A and B are $m \times n$ matrices. Then $(A+B)^t = A^t + B^t$. □

Proof The statement to be proved is an equality of matrices, so we work entry-by-entry. Think carefully about the objects involved here, and the many uses of the plus sign.

$$\begin{aligned}(A+B)^t{}_{ij} &= A + B_{ji} \\ &= A_{ji} + B_{ji} \\ &= A^t{}_{ij} + B^t{}_{ij} \\ &= A^t + B^t{}_{ij}\end{aligned}$$

Since the matrices $(A+B)^t$ and $A^t + B^t$ agree at each entry, it tells us the two matrices are equal. ■

Theorem 52 (Transpose and Matrix Scalar Multiplication) Suppose that $\alpha \in \mathbb{R}$ and A is an $m \times n$ matrix. Then $(\alpha A)^t = \alpha A^t$. □

Proof The statement to be proved is an equality of matrices, so we work entry-by-entry. Think carefully about the objects involved here, the many uses of juxtaposition.

$$\begin{aligned}(\alpha A)^t{}_{ij} &= \alpha A_{ji} \\ &= \alpha A_{ji} \\ &= \alpha A^t{}_{ij} \\ &= \alpha A^t{}_{ij}\end{aligned}$$

Since the matrices $(\alpha A)^t$ and αA^t agree at each entry, it tells us the two matrices are equal. ■

Theorem 53 (Transpose of a Transpose) Suppose that A is an $m \times n$ matrix. Then $(A^t)^t = A$. □

Proof We again want to prove an equality of matrices, so we work entry-by-entry.

$$\begin{aligned}\left(A^t\right)^t{}_{ij} &= A^t{}_{ji} \\ &= A_{ij}\end{aligned}$$

■

Definition 54 (Matrix-Vector Product) Suppose A is an $m \times n$ matrix with columns $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$, $\ldots$, $\boldsymbol{A}_n$ and $\boldsymbol{u}$ is a vector of size n. Then the **matrix-vector product** of A with $\boldsymbol{u}$ is the linear combination

$$A\boldsymbol{u} = \boldsymbol{u}_1\boldsymbol{A}_1 + \boldsymbol{u}_2\boldsymbol{A}_2 + \boldsymbol{u}_3\boldsymbol{A}_3 + \cdots + \boldsymbol{u}_n\boldsymbol{A}_n \qquad \triangle$$

So, the matrix-vector product is yet another version of "multiplication," at least in the sense that we have yet again overloaded juxtaposition of two symbols as our notation. Remember your objects, an $m \times n$ matrix times a vector of size n will create a vector of size m. So if A is rectangular, then the size of the vector changes. With all the linear combinations we have performed so far, this computation should now seem second nature.

Example 55 (A matrix times a vector) Consider

$$A = \begin{bmatrix} 1 & 4 & 2 & 3 & 4 \\ -3 & 2 & 0 & 1 & -2 \\ 1 & 6 & -3 & -1 & 5 \end{bmatrix} \qquad \boldsymbol{u} = \begin{bmatrix} 2 \\ 1 \\ -2 \\ 3 \\ -1 \end{bmatrix}$$

Then

$$A\boldsymbol{u} = 2\begin{bmatrix} 1 \\ -3 \\ 1 \end{bmatrix} + 1\begin{bmatrix} 4 \\ 2 \\ 6 \end{bmatrix} + (-2)\begin{bmatrix} 2 \\ 0 \\ -3 \end{bmatrix} + 3\begin{bmatrix} 3 \\ 1 \\ -1 \end{bmatrix} + (-1)\begin{bmatrix} 4 \\ -2 \\ 5 \end{bmatrix} = \begin{bmatrix} 7 \\ 1 \\ 6 \end{bmatrix}. \qquad \boxtimes$$

This definition now makes it possible to represent systems of linear equations compactly in terms of an operation.

Theorem 56 (Systems of Linear Equations as Matrix Multiplication) Solutions to the linear system $\mathcal{LS}(A, \boldsymbol{b})$ are the solutions for $\boldsymbol{x}$ in the vector equation $A\boldsymbol{x} = \boldsymbol{b}$. $\square$

Proof This theorem says that two sets (of solutions) are equal. So we need to show that one set of solutions is a subset of the other, and vice versa. Let $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$, $\ldots$, $\boldsymbol{A}_n$ be the columns of A. Both of these set inclusions then follow from the following chain of equivalences,

$$\begin{aligned} &\boldsymbol{x} \text{ is a solution to } \mathcal{LS}(A, \boldsymbol{b}) \\ &\iff \boldsymbol{x}_1\boldsymbol{A}_1 + \boldsymbol{x}_2\boldsymbol{A}_2 + \boldsymbol{x}_3\boldsymbol{A}_3 + \cdots + \boldsymbol{x}_n\boldsymbol{A}_n = \boldsymbol{b} \\ &\iff \boldsymbol{x} \text{ is a solution to } A\boldsymbol{x} = \boldsymbol{b} \end{aligned}$$

■

Example 57 (Matrix notation for systems of linear equations) Consider the following system of linear equations

$$\begin{aligned} 2x_1 + 4x_2 - 3x_3 + 5x_4 + x_5 &= 9 \\ 3x_1 + x_2 + x_4 - 3x_5 &= 0 \\ -2x_1 + 7x_2 - 5x_3 + 2x_4 + 2x_5 &= -3 \end{aligned}$$

has coefficient matrix

$$A = \begin{bmatrix} 2 & 4 & -3 & 5 & 1 \\ 3 & 1 & 0 & 1 & -3 \\ -2 & 7 & -5 & 2 & 2 \end{bmatrix}$$

and vector of constants

$$\boldsymbol{b} = \begin{bmatrix} 9 \\ 0 \\ -3 \end{bmatrix}$$

and so will be described compactly by the vector equation $A\boldsymbol{x} = \boldsymbol{b}$. ⊠

Definition 58 (Nonsingular Matrix) Suppose A is a square matrix. Suppose further that the solution set to the homogeneous linear system of equations $\mathcal{LS}(A, \mathbf{0})$ is $\{\mathbf{0}\}$, i.e. the system has *only* the trivial solution. Then we say that A is a **nonsingular** matrix. Otherwise we say A is a **singular** matrix. △

Theorem 59 (Equal Matrices and Matrix-Vector Products) Suppose that A and B are $m \times n$ matrices such that $A\boldsymbol{x} = B\boldsymbol{x}$ for every $\boldsymbol{x} \in \mathbb{R}^n$. Then $A = B$. □

Proof Since $A\boldsymbol{x} = B\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^n$, choose $\boldsymbol{x}$ to be a vector of all zeros, with a lone 1 in the i-th slot. Then

$$\begin{aligned} A\boldsymbol{x} &= [\boldsymbol{A}_1|\boldsymbol{A}_2|\boldsymbol{A}_3|\dots|\boldsymbol{A}_n] \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &= 0\boldsymbol{A}_1 + 0\boldsymbol{A}_2 + 0\boldsymbol{A}_3 + \cdots + 0\boldsymbol{A}_{i-1} + 1\boldsymbol{A}_i + 0\boldsymbol{A}_{i+1} + \cdots + 0\boldsymbol{A}_n \\ &= \boldsymbol{A}_i \end{aligned}$$

Similarly, $B\boldsymbol{x} = \boldsymbol{B}_i$, so $\boldsymbol{A}_i = \boldsymbol{B}_i$, $1 \leq i \leq n$ and so all the columns of A and B are equal. Then our definition of column vector equality establishes that the individual entries of A and B in each column are equal. So the matrices A and B are equal. ■

Definition 60 (Matrix Multiplication) Suppose A is an $m \times n$ matrix and B is an $n \times p$ matrix with columns $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, $\boldsymbol{B}_3$, ..., $\boldsymbol{B}_p$. Then the **matrix product** of A with B is the $m \times p$ matrix where column i is the matrix-vector product $A\boldsymbol{B}_i$. Symbolically,

$$AB = A\left[\boldsymbol{B}_1|\boldsymbol{B}_2|\boldsymbol{B}_3|\ldots|\boldsymbol{B}_p\right] = \left[A\boldsymbol{B}_1|A\boldsymbol{B}_2|A\boldsymbol{B}_3|\ldots|A\boldsymbol{B}_p\right]. \qquad \triangle$$

Example 61 (Product of two matrices) Set

$$A = \begin{bmatrix} 1 & 2 & -1 & 4 & 6 \\ 0 & -4 & 1 & 2 & 3 \\ -5 & 1 & 2 & -3 & 4 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 6 & 2 & 1 \\ -1 & 4 & 3 & 2 \\ 1 & 1 & 2 & 3 \\ 6 & 4 & -1 & 2 \\ 1 & -2 & 3 & 0 \end{bmatrix}$$

Then

$$AB = \left[\begin{array}{c|c|c|c} A\begin{bmatrix} 1 \\ -1 \\ 1 \\ 6 \\ 1 \end{bmatrix} & A\begin{bmatrix} 6 \\ 4 \\ 1 \\ 4 \\ -2 \end{bmatrix} & A\begin{bmatrix} 2 \\ 3 \\ 2 \\ -1 \\ 3 \end{bmatrix} & A\begin{bmatrix} 1 \\ 2 \\ 3 \\ 2 \\ 0 \end{bmatrix} \end{array}\right] = \begin{bmatrix} 28 & 17 & 20 & 10 \\ 20 & -13 & -3 & -1 \\ -18 & -44 & 12 & -3 \end{bmatrix}.$$

⊠

Is this the definition of matrix multiplication you expected? Perhaps our previous operations for matrices caused you to think that we might multiply two matrices of the *same* size, *entry-by-entry*? Notice that our current definition uses matrices of different sizes (though the number of columns in the first must equal the number of rows in the second), and the result is of a third size. Notice too in the previous example that we cannot even consider the product BA, since the sizes of the two matrices in this order aren't right.

But it gets weirder than that. Many of your old ideas about "multiplication" won't apply to matrix multiplication, but some still will. So make no assumptions, and don't do anything until you have a theorem that says you can. Even if the sizes are right, matrix multiplication is not commutative — order matters.

Example 62 (Matrix Multiplication is not commutative) Set

$$A = \begin{bmatrix} 1 & 3 \\ -1 & 2 \end{bmatrix} \qquad\qquad B = \begin{bmatrix} 4 & 0 \\ 5 & 1 \end{bmatrix}.$$

Then we have two square, 2×2 matrices, so we are allow to multiply them in either order. We find

$$AB = \begin{bmatrix} 19 & 3 \\ 6 & 2 \end{bmatrix} \qquad\qquad BA = \begin{bmatrix} 4 & 12 \\ 4 & 17 \end{bmatrix} \qquad \boxtimes$$

and $AB \neq BA$. Not even close. It should not be hard for you to construct other pairs of matrices that do not commute (try a couple of 3×3's). Can you find a pair of non-identical matrices that *do* commute?

Theorem 63 (Entries of Matrix Products) Suppose A is an $m \times n$ matrix and $B =$ is an $n \times p$ matrix. Then for $1 \leq i \leq m$, $1 \leq j \leq p$, the individual entries of AB are given by

$$\begin{aligned} AB_{ij} &= A_{i1}B_{1j} + A_{i2}B_{2j} + A_{i3}B_{3j} + \cdots + A_{in}B_{nj} \\ &= \sum_{k=1}^{n} A_{ik}B_{kj} \end{aligned} \qquad \square$$

Proof Denote the columns of A as the vectors $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$, $\ldots$, $\boldsymbol{A}_n$ and the columns of B as the vectors $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, $\boldsymbol{B}_3$, $\ldots$, $\boldsymbol{B}_p$. Then for $1 \leq i \leq m$, $1 \leq j \leq p$,

$$\begin{aligned} AB_{ij} &= A\boldsymbol{B}_{j_i} \\ &= \boldsymbol{B}_{j_1}\boldsymbol{A}_1 + \boldsymbol{B}_{j_2}\boldsymbol{A}_2 + \boldsymbol{B}_{j_3}\boldsymbol{A}_3 + \cdots + \boldsymbol{B}_{j_n}\boldsymbol{A}_{n_i} \\ &= \boldsymbol{B}_{j_1}\boldsymbol{A}_{1_i} + \boldsymbol{B}_{j_2}\boldsymbol{A}_{2_i} + \boldsymbol{B}_{j_3}\boldsymbol{A}_{3_i} + \cdots + \boldsymbol{B}_{j_n}\boldsymbol{A}_{n_i} \\ &= \boldsymbol{B}_{j_1}\boldsymbol{A}_{1i} + \boldsymbol{B}_{j_2}\boldsymbol{A}_{2i} + \boldsymbol{B}_{j_3}\boldsymbol{A}_{3i} + \cdots + \boldsymbol{B}_{j_n}\boldsymbol{A}_{ni} \\ &= B_{1j}A_{i1} + B_{2j}A_{i2} + B_{3j}A_{i3} + \cdots + B_{nj}A_{in} \\ &= A_{i1}B_{1j} + A_{i2}B_{2j} + A_{i3}B_{3j} + \cdots + A_{in}B_{nj} \\ &= \sum_{k=1}^{n} A_{ik}B_{kj} \end{aligned} \qquad \blacksquare$$

Example 64 (Product of two matrices, entry-by-entry) Consider again the two matrices

$$A = \begin{bmatrix} 1 & 2 & -1 & 4 & 6 \\ 0 & -4 & 1 & 2 & 3 \\ -5 & 1 & 2 & -3 & 4 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 6 & 2 & 1 \\ -1 & 4 & 3 & 2 \\ 1 & 1 & 2 & 3 \\ 6 & 4 & -1 & 2 \\ 1 & -2 & 3 & 0 \end{bmatrix}$$

Then suppose we just wanted the entry of AB in the second row, third column:

$$\begin{aligned} AB_{23} =& A_{21}B_{13} + A_{22}B_{23} + A_{23}B_{33} + A_{24}B_{43} + A_{25}B_{53} \\ =& (0)(2) + (-4)(3) + (1)(2) + (2)(-1) + (3)(3) = -3 \end{aligned}$$

Notice how there are 5 terms in the sum, since 5 is the common dimension of the two matrices (column count for A, row count for B).

The entry of third row, first column:

$$\begin{aligned} AB_{31} =& A_{31}B_{11} + A_{32}B_{21} + A_{33}B_{31} + A_{34}B_{41} + A_{35}B_{51} \\ =& (-5)(1) + (1)(-1) + (2)(1) + (-3)(6) + (4)(1) = -18 \end{aligned}$$

⊠

To get some more practice on your own, complete the computation of the other 10 entries of this product. Construct some other pairs of matrices (of compatible sizes) and compute their product two ways.

Theorem 65 (Matrix Multiplication and the Zero Matrix) Suppose A is an $m \times n$ matrix. Then 1. $A\mathcal{O}_{n\times p} = \mathcal{O}_{m\times p}$ 2. $\mathcal{O}_{p\times m}A = \mathcal{O}_{p\times n}$ □

Proof We'll prove (1) and leave (2) to you. Entry-by-entry,

$$\begin{aligned} A\mathcal{O}_{n\times p_{ij}} &= \sum_{k=1}^{n} A_{ik}\mathcal{O}_{n\times p_{kj}} \\ &= \sum_{k=1}^{n} A_{ik}0 \\ &= \sum_{k=1}^{n} 0 = 0. \end{aligned}$$

So every entry of the product is the scalar zero, i.e. the result is the zero matrix. ■

Theorem 66 (Matrix Multiplication and Identity Matrix) Suppose A is an $m \times n$ matrix. Then

1. $AI_n = A$
2. $I_m A = A$ □

Proof Again, we'll prove (1) and leave (2) to you. Entry-by-entry,

$$\begin{aligned}
AI_{nij} =& \sum_{k=1}^{n} A_{ik} I_{nkj} \\
=& A_{ij} I_{njj} + \sum_{k=1, k\neq j}^{n} A_{ik} I_{nkj} \\
=& A_{ij}(1) + \sum_{k=1, k\neq j}^{n} A_{ik}(0) \\
=& A_{ij} + \sum_{k=1, k\neq j}^{n} 0 \\
=& A_{ij}
\end{aligned}$$

■

So the matrices A and AI_n are equal, entry-by-entry, and by the definition of matrix equality we can say they are equal matrices.

It is this theorem that gives the identity matrix its name. It is a matrix that behaves with matrix multiplication like the scalar 1 does with scalar multiplication. To multiply by the identity matrix is to have no effect on the other matrix.

Theorem 67 (Matrix Multiplication Distributes Across Addition) Suppose A is an $m \times n$ matrix and B and C are $n \times p$ matrices and D is a $p \times s$ matrix. Then 1. $A(B+C) = AB + AC$ 2. $(B+C)D = BD + CD$□

Proof We'll do (1), you do (2). Entry-by-entry,

$$\begin{aligned}
A(B+C)_{ij} =& \sum_{k=1}^{n} A_{ik}B + C_{kj} \\
=& \sum_{k=1}^{n} A_{ik}(B_{kj} + C_{kj}) \\
=& \sum_{k=1}^{n} A_{ik}B_{kj} + A_{ik}C_{kj} \\
=& \sum_{k=1}^{n} A_{ik}B_{kj} + \sum_{k=1}^{n} A_{ik}C_{kj} \\
=& AB_{ij} + AC_{ij} \\
=& AB + AC_{ij}
\end{aligned}$$

So the matrices $A(B+C)$ and $AB+AC$ are equal, entry-by-entry, and by the definition of matrix equality we can say they are equal matrices. ■

Theorem 68 (Matrix Multiplication and Scalar Matrix Multiplication) Suppose A is an $m \times n$ matrix and B is an $n \times p$ matrix. Let α be a scalar. Then $\alpha(AB) = (\alpha A)B = A(\alpha B)$. □

Proof These are equalities of matrices. We'll do the first one, the second is similar and will be good practice for you.

$$\begin{aligned}
\alpha(AB)_{ij} =& \alpha AB_{ij} \\
=& \alpha \sum_{k=1}^{n} A_{ik}B_{kj} \\
=& \sum_{k=1}^{n} \alpha A_{ik}B_{kj} \\
=& \sum_{k=1}^{n} \alpha A_{ik}B_{kj} \\
=& (\alpha A)B_{ij}
\end{aligned}$$

So the matrices $\alpha(AB)$ and $(\alpha A)B$ are equal, entry-by-entry, and by the definition of matrix equality we can say they are equal matrices. ■

Theorem 69 (Matrix Multiplication is Associative) Suppose A is an $m \times n$ matrix, B is an $n \times p$ matrix and D is a $p \times s$ matrix. Then $A(BD) = (AB)D$. □

Proof A matrix equality, so we'll go entry-by-entry, no surprise there.

$$\begin{aligned}
A(BD)_{ij} &= \sum_{k=1}^{n} A_{ik} BD_{kj} \\
&= \sum_{k=1}^{n} A_{ik} \left(\sum_{\ell=1}^{p} B_{k\ell} D_{\ell j} \right) \\
&= \sum_{k=1}^{n} \sum_{\ell=1}^{p} A_{ik} B_{k\ell} D_{\ell j}
\end{aligned}$$

We can switch the order of the summation since these are finite sums,

$$= \sum_{\ell=1}^{p} \sum_{k=1}^{n} A_{ik} B_{k\ell} D_{\ell j}$$

As $D_{\ell j}$ does not depend on the index k, we can factor it out of the inner sum,

$$\begin{aligned}
&= \sum_{\ell=1}^{p} D_{\ell j} \left(\sum_{k=1}^{n} A_{ik} B_{k\ell} \right) \\
&= \sum_{\ell=1}^{p} D_{\ell j} AB_{i\ell} \\
&= \sum_{\ell=1}^{p} AB_{i\ell} D_{\ell j} \\
&= (AB)D_{ij}
\end{aligned}$$

So the matrices $(AB)D$ and $A(BD)$ are equal, entry-by-entry, and by the definition of matrix equality we can say they are equal matrices. ■

Theorem 70 (Matrix Multiplication and Inner Products) If we consider the vectors $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{R}^m$ as $m \times 1$ matrices then

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^t \boldsymbol{v}$$ □

Proof

$$\begin{aligned}\langle \boldsymbol{u},\, \boldsymbol{v}\rangle &= \sum_{k=1}^{m} \boldsymbol{u}_k \boldsymbol{v}_k \\ &= \sum_{k=1}^{m} \boldsymbol{u}_{k1} \boldsymbol{v}_{k1} \\ &= \sum_{k=1}^{m} \boldsymbol{u}^t{}_{1k} \boldsymbol{v}_{k1} \\ &= \sum_{k=1}^{m} \boldsymbol{u}^t{}_{1k} \boldsymbol{v}_{k1} \\ &= \boldsymbol{u}^t \boldsymbol{v}_{11}\end{aligned}$$

To finish we just blur the distinction between a 1×1 matrix ($\boldsymbol{u}^t\boldsymbol{v}$) and its lone entry. ■

Theorem 71 (Matrix Multiplication and Transposes) Suppose A is an $m \times n$ matrix and B is an $n \times p$ matrix. Then $(AB)^t = B^t A^t$. □

Proof This theorem may be surprising but if we check the sizes of the matrices involved, then maybe it will not seem so far-fetched. First, AB has size $m \times p$, so its transpose has size $p \times m$. The product of B^t with A^t is a $p \times n$ matrix times an $n \times m$ matrix, also resulting in a $p \times m$ matrix. So at least our objects are compatible for equality (and would not be, in general, if we didn't reverse the order of the operation).

Here we go again, entry-by-entry,

$$\begin{aligned}(AB)^t{}_{ij} &= AB_{ji} \\ &= \sum_{k=1}^{n} A_{jk} B_{ki} \\ &= \sum_{k=1}^{n} B_{ki} A_{jk} \\ &= \sum_{k=1}^{n} B^t{}_{ik} A^t{}_{kj} \\ &= B^t A^t{}_{ij}\end{aligned}$$

So the matrices $(AB)^t$ and B^tA^t are equal, entry-by-entry, and by the definition of matrix equality we can say they are equal matrices. ■

This theorem seems odd at first glance, since we have to switch the order of A and B. But if we simply consider the sizes of the matrices involved, we can see that the switch is necessary for this reason alone. That the individual entries of the products then come along to be equal is a bonus.

Notice how none of these proofs above relied on writing out huge general matrices with lots of ellipses ("...") and trying to formulate the equalities a whole matrix at a time. This messy business is a "proof technique" to be avoided at all costs.

These theorems give you the "rules" for how matrices interact with the various operations we have defined. Use them and use them often. But don't try to do anything with a matrix that you don't have a rule for. Together, we would informally call all these operations, and the attendant theorems, "the algebra of matrices." Notice, too, that every column vector is just a $n \times 1$ matrix, so these theorems apply to column vectors also. Finally, these results may make us feel that the definition of matrix multiplication is not so unnatural.

Definition 72 (Matrix Inverse) Suppose A and B are square matrices of size n such that $AB = I_n$ and $BA = I_n$. Then A is **invertible** and B is the **inverse** of A. In this situation, we write $B = A^{-1}$. △

- Notice that if B is the inverse of A, then we can just as easily say A is the inverse of B, or A and B are inverses of each other.
- Not every square matrix has an inverse.

We will have occasion in this subsection (and later) to reference the following frequently used vectors, so we will make a useful definition now.

Definition 73 (Standard Unit Vectors) Let $\boldsymbol{e}_j \in \mathbb{R}^m$ denote the column vector that is column j of the $m \times m$ identity matrix I_m. Then the set

$$\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \ldots, \boldsymbol{e}_m\} = \{\boldsymbol{e}_j \mid 1 \leq j \leq m\}$$

is the set of **standard unit vectors** in $\mathbb{R}^m$. △

We will make reference to these vectors often. Notice that $\boldsymbol{e}_j$ is a column vector full of zeros, with a lone 1 in the j-th position, so an alternate definition is

$$\boldsymbol{e}_{ji} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

Theorem 74 (Two-by-Two Matrix Inverse) Suppose

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Then A is invertible if and only if $ad - bc \neq 0$. When A is invertible, we have

$$A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \qquad \square$$

Proof ($\Leftarrow$) If $ad - bc \neq 0$ then the displayed formula is legitimate (we are not dividing by zero), and it is a simple matter to actually check that $A^{-1}A = AA^{-1} = I_2$.

($\Rightarrow$) Assume that A is invertible, and proceed with a proof by contradiction, by assuming also that $ad - bc = 0$. This means that $ad = bc$. Let

$$B = \begin{bmatrix} e & f \\ g & h \end{bmatrix}$$

be a putative inverse of A. This means that

$$I_2 = AB = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}$$

Working on the matrices on both ends of this equation, we will multiply the top row by c and the bottom row by a.

$$\begin{bmatrix} c & 0 \\ 0 & a \end{bmatrix} = \begin{bmatrix} ace+bcg & acf+bch \\ ace+adg & acf+adh \end{bmatrix}$$

We are assuming that $ad = bc$, so we can replace two occurences of ad by bc in the bottom row of the right matrix.

$$\begin{bmatrix} c & 0 \\ 0 & a \end{bmatrix} = \begin{bmatrix} ace+bcg & acf+bch \\ ace+bcg & acf+bch \end{bmatrix}$$

The matrix on the right now has two rows that are identical, and therefore the same must be true of the matrix on the left. Given the form of the matrix on the left, identical rows implies that $a = 0$ and $c = 0$.

With this information, the product AB becomes

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2 = AB = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix} = \begin{bmatrix} bg & bh \\ dg & dh \end{bmatrix}$$

So $bg = dh = 1$ and thus b, g, d, h are all nonzero. But then bh and dg (the "other corners") must also be nonzero, so this is (finally) a contradiction. So our assumption was false and we see that $ad - bc \neq 0$ whenever A has an inverse. ■

Example 75 (Computing a Matrix Inverse) Let A be a square matrix. For its inverse, we desire a matrix B so that $AB = I_5$. Emphasizing the structure of the columns and employing the definition of matrix multiplication,

$$\begin{aligned} AB &= I_5 \\ A[\boldsymbol{B}_1|\boldsymbol{B}_2|\boldsymbol{B}_3|\boldsymbol{B}_4|\boldsymbol{B}_5] &= [\boldsymbol{e}_1|\boldsymbol{e}_2|\boldsymbol{e}_3|\boldsymbol{e}_4|\boldsymbol{e}_5] \\ [A\boldsymbol{B}_1|A\boldsymbol{B}_2|A\boldsymbol{B}_3|A\boldsymbol{B}_4|A\boldsymbol{B}_5] &= [\boldsymbol{e}_1|\boldsymbol{e}_2|\boldsymbol{e}_3|\boldsymbol{e}_4|\boldsymbol{e}_5]. \end{aligned}$$

Equating the matrices column-by-column we have

$$A\boldsymbol{B}_1 = \boldsymbol{e}_1 \qquad A\boldsymbol{B}_2 = \boldsymbol{e}_2 \qquad A\boldsymbol{B}_3 = \boldsymbol{e}_3 \qquad A\boldsymbol{B}_4 = \boldsymbol{e}_4 \qquad A\boldsymbol{B}_5 = \boldsymbol{e}_5.$$

Since the matrix B is what we are trying to compute, we can view each column, $\boldsymbol{B}_i$, as a column vector of unknowns. Then we have five systems of equations to solve, each with 5 equations in 5 variables. Notice that all 5 of these systems have the same coefficient matrix. ⊠

Theorem 76 (Matrix Inverse is Unique) Suppose the square matrix A has an inverse. Then A^{-1} is unique. □

Proof We will assume that A has two inverses. The hypothesis tells there is at least one. Suppose then that B and C are both inverses for A. Then,

$$\begin{aligned} B &= BI_n \\ &= B(AC) \\ &= (BA)C \\ &= I_nC \\ &= C \end{aligned}$$

So we conclude that B and C are the same, and cannot be different. So any matrix that acts like *an* inverse, must be *the* inverse. ■

When most of us dress in the morning, we put on our socks first, followed by our shoes. In the evening we must then first remove our shoes, followed by our socks. Try to connect the conclusion of the following theorem with this everyday example.

Theorem 77 (Socks and Shoes) Suppose A and B are invertible matrices of size n. Then $(AB)^{-1} = B^{-1}A^{-1}$ and AB is an invertible matrix. □

Proof At the risk of carrying our everyday analogies too far, the proof of this theorem is quite easy when we compare it to the workings of a dating service. We have a statement about the inverse of the matrix AB, which for all we know right now might not even exist. Suppose AB was to sign up for a dating service with two requirements for a compatible date. Upon multiplication on the left, and on the right, the result should be the identity matrix. In other words, AB's ideal date would be its inverse.

Now along comes the matrix $B^{-1}A^{-1}$ (which we know exists because our hypothesis says both A and B are invertible and we can form the product of these two matrices), also looking for a date. Let's see if $B^{-1}A^{-1}$ is a good match for AB. First they meet at a non-committal neutral location, say a coffee shop, for quiet conversation:

$$\begin{aligned}(B^{-1}A^{-1})(AB) &= B^{-1}(A^{-1}A)B\\ &= B^{-1}I_nB\\ &= B^{-1}B\\ &= I_n\end{aligned}$$

The first date having gone smoothly, a second, more serious, date is arranged, say dinner and a show:

$$\begin{aligned}(AB)(B^{-1}A^{-1}) &= A(BB^{-1})A^{-1}\\ &= AI_nA^{-1}\\ &= AA^{-1}\\ &= I_n\end{aligned}$$

So the matrix $B^{-1}A^{-1}$ has met all of the requirements to be AB's inverse (date) and with the ensuing marriage proposal we can announce that $(AB)^{-1} = B^{-1}A^{-1}$. ■

Theorem 78 (Matrix Inverse of a Matrix Inverse) Suppose A is an invertible matrix. Then A^{-1} is invertible and $(A^{-1})^{-1} = A$. □

Proof We examine if A is a suitable inverse for A^{-1} (by definition, the opposite is true).

$$AA^{-1} = I_n$$

and

$$A^{-1}A = I_n$$

The matrix A has met all the requirements to be the inverse of A^{-1}, and so is invertible and we can write $A = (A^{-1})^{-1}$. ■

Theorem 79 (Matrix Inverse of a Transpose) Suppose A is an invertible matrix. Then A^t is invertible and $(A^t)^{-1} = (A^{-1})^t$. □

Proof We see if $(A^{-1})^t$ is a suitable inverse for A^t. Apply the theorem about Matrix multiplication and Transpose to see that

$$\begin{aligned}(A^{-1})^t A^t &= (AA^{-1})^t \\ &= {I_n}^t \\ &= I_n && I_n \text{ is symmetric}\end{aligned}$$

and

$$\begin{aligned}A^t(A^{-1})^t &= (A^{-1}A)^t \\ &= {I_n}^t \\ &= I_n && I_n \text{ is symmetric}\end{aligned}$$

The matrix $(A^{-1})^t$ has met all the requirements to be the inverse of A^t, and so is invertible and we can write $(A^t)^{-1} = (A^{-1})^t$. ■

Theorem 80 (Matrix Inverse of a Scalar Multiple) Suppose A is an invertible matrix and α is a nonzero scalar. Then $(\alpha A)^{-1} = \frac{1}{\alpha}A^{-1}$ and αA is invertible. □

Proof We see if $\frac{1}{\alpha}A^{-1}$ is a suitable inverse for αA.

$$\begin{aligned}\left(\frac{1}{\alpha}A^{-1}\right)(\alpha A) &= \left(\frac{1}{\alpha}\alpha\right)\left(AA^{-1}\right) \\ &= 1I_n && \text{Scalar multiplicative inverses} \\ &= I_n\end{aligned}$$

and

$$\begin{aligned}(\alpha A)\left(\frac{1}{\alpha}A^{-1}\right) &= \left(\alpha\frac{1}{\alpha}\right)\left(A^{-1}A\right) \\ &= 1I_n && \text{Scalar multiplicative inverses} \\ &= I_n\end{aligned}$$

The matrix $\frac{1}{\alpha}A^{-1}$ has met all the requirements to be the inverse of αA, so we can write $(\alpha A)^{-1} = \frac{1}{\alpha}A^{-1}$. ■

Theorem 81 (Inverse of a Partitioned Matrix) For partitioned nonsingular matrix $A : n \times n$

$$A = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$$

where $E : n_1 \times n_1$, $F : n_1 \times n_2$, $G : n_2 \times n_1$ and $H : n_2 \times n_2$, $(n = n_1 + n_2)$ are such that E and $H - GE^{-1}F$ are nonsingular, the partitioned inverse is given by

$$A^{-1} = \begin{bmatrix} E^{-1}(I + FD^{-1}GG^{-1}) & -E^{-1}FD^{-1} \\ -D^{-1}GE^{-1} & D^{-1} \end{bmatrix} \qquad \square$$

Proof Check that the product of A and A^{-1} reduces to the identity matrix.■

-3.10.1 Trace

Definition 82 Let A_{ii} (for all $1 \leq i \leq n$) be the elements on the main diagonal of a square matrix $A : n \times n$. Then the **trace** of A is definedd as the sum

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

Theorem 83 Let A and B be square $n \times n$ matrices. Then we have

$$\mathrm{tr}(A \pm B) = \mathrm{tr}(A) \pm \mathrm{tr}(B) \qquad \square$$

Proof By definition of sum between two matrices we have: $C = A + B$ means $C_{ij} = A_{ij} + B_{ij}$ hence

$$\begin{aligned} \mathrm{tr}(C) &= \sum_{i=1}^{n} C_{ii} \\ &= \sum_{i=1}^{n} A_{ii} + B_{ii} \\ &= \sum_{i=1}^{n} A_{ii} + \sum_{i=1}^{n} B_{ii} \\ &\mathrm{tr}(A) + \mathrm{tr}(B) \end{aligned}$$

The same holds for $D = A - B$. ■

Theorem 84 Let A be square $n \times n$ matrix. Then we have

$$\mathrm{tr}(A^t) = \mathrm{tr}(A) \qquad \square$$

Proof The result follows from the fact ${A^t}_{ii} = A_{ii}$ for all $1 \leq i \leq n$. ■

Theorem 85 Let A be square $n \times n$ matrix and c be a scalar factor. Then we have

$$\mathrm{tr}(cA) = c\,\mathrm{tr}(A) \qquad \square$$

Proof By definition of the Matrix Scalar multiplication the elements of the main diagonal of cA are on the form cA_{ii} so that

$$\mathrm{tr}(cA) = \sum_{i=1}^{n} cA_{ii} = c\sum_{i=1}^{n} A_{ii} = c\,\mathrm{tr}(A) \qquad \blacksquare$$

Theorem 86 Let $A : m \times n$ and $B : n \times m$ matrices. Then we have

$$\mathrm{tr}(AB) = \mathrm{tr}(BA) \qquad \square$$

Proof Let $C = AB$ and note that $C_{ii} = \langle \boldsymbol{A}_i, \boldsymbol{B^t}_i \rangle = \sum_{k=1}^{n} A_{ik}B_{ki}$ so that

$$\mathrm{tr}(C) = \sum_{i=1}^{m}\sum_{k=1}^{n} A_{ik}B_{ki}$$

In the same way let $D = BA$ and

$$\mathrm{tr}(D) = \sum_{i=1}^{n}\sum_{k=1}^{m} B_{ki}A_{ik}$$

But than $\mathrm{tr}(C) = \mathrm{tr}(D)$ since the commutative property of multiplication and the possibility fo interchange between summation operators. ■

Theorem 87 Let A be square $n \times n$ matrix. Then we have

$$\mathrm{tr}(AA^t) = \mathrm{tr}(A^tA) = \sum_{i=1}^{n}\sum_{j=1}^{n} {A_{ij}}^2 \qquad \square$$

Proof The first equality follows from the previous theorem. The second equality follows from the Entries of Matrix Products theorem since ${AA^t}_{ii} = \sum_{k=1}^{n} A_{ik}A_{ki}$. ■

Theorem 88 The squared norm of the $\boldsymbol{a}$ is

$$\|\boldsymbol{a}\|^2 = \langle \boldsymbol{a}, \boldsymbol{a} \rangle = \sum_{i=1}^{n} {\boldsymbol{a}_i}^2 = \mathrm{tr}(\boldsymbol{a}\boldsymbol{a}^t) \qquad \square$$

Proof Since the Inner Products and Norms Theorem we have the first equality. The second follows from the definition of inner product and the last follows from the previous theorem. ■

-3.10.2 Determinant of a Matrix

The definition of the determinant function is **recursive**, that is, the determinant of a large matrix is defined in terms of the determinant of smaller matrices. To this end, we will make a few definitions.

Definition 89 (SubMatrix) Suppose that A is an $m \times n$ matrix. Then the **submatrix** A_{ij} is the $(m-1) \times (n-1)$ matrix obtained from A by removing row i and column j. △

Example 90 (Some submatrices) For the matrix

$$A = \begin{bmatrix} 1 & -2 & 3 & 9 \\ 4 & -2 & 0 & 1 \\ 3 & 5 & 2 & 1 \end{bmatrix}$$

we have the submatrices

$$A_{23} = \begin{bmatrix} 1 & -2 & 9 \\ 3 & 5 & 1 \end{bmatrix} \qquad A_{31} = \begin{bmatrix} -2 & 3 & 9 \\ -2 & 0 & 1 \end{bmatrix} \qquad \boxtimes$$

Definition 91 (Determinant of a Matrix) Suppose A is a square matrix. Then its **determinant**, $\det(A) = |A|$, is an element of $\mathbb{C}$ defined recursively by:[6pt] If $A = \begin{bmatrix} a \end{bmatrix}$ is a 1×1 matrix, then $\det(A) = a$.[6pt] If A is a matrix of size n with $n \geq 2$, then

$$\det(A) = A_{11} \det(A_{11}) - A_{12} \det(A_{12}) + A_{13} \det(A_{13}) - \cdots + (-1)^{n+1} A_{1n} \det(A_{1n})$$

△

So to compute the determinant of a 5×5 matrix we must build 5 submatrices, each of size 4. To compute the determinants of each the 4×4 matrices we need to create 4 submatrices each, these now of size 3 and so on. To compute the determinant of a 10×10 matrix would require computing the determinant

of $10! = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 = 3,628,800$ 1×1 matrices. Fortunately there are better ways. However this does suggest an excellent computer programming exercise to write a recursive procedure to compute a determinant.

Let's compute the determinant of a reasonable sized matrix by hand.

Example 92 (Determinant of a 3×3 matrix) Suppose that we have the 3×3 matrix

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

Then

$$\begin{aligned} \det(A) = |A| &= \begin{vmatrix} 3 & 2 & -1 \\ 4 & 1 & 6 \\ -3 & -1 & 2 \end{vmatrix} \\ &= 3\begin{vmatrix} 1 & 6 \\ -1 & 2 \end{vmatrix} - 2\begin{vmatrix} 4 & 6 \\ -3 & 2 \end{vmatrix} + (-1)\begin{vmatrix} 4 & 1 \\ -3 & -1 \end{vmatrix} \\ &= 3\left(1\left|2\right| - 6\left|-1\right|\right) - 2\left(4\left|2\right| - 6\left|-3\right|\right) - \left(4\left|-1\right| - 1\left|-3\right|\right) \\ &= 3\left(1(2) - 6(-1)\right) - 2\left(4(2) - 6(-3)\right) - \left(4(-1) - 1(-3)\right) \\ &= 24 - 52 + 1 \\ &= -27 \end{aligned}$$ ⊠

In practice it is a bit silly to decompose a 2×2 matrix down into a couple of 1×1 matrices and then compute the exceedingly easy determinant of these puny matrices. So here is a simple theorem.

Theorem 93 (Determinant of Matrices of Size Two) Suppose that $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Then $\det(A) = ad - bc$ □

Proof

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = a\left|d\right| - b\left|c\right| = ad - bc$$ ■

Do you recall seeing the expression $ad - bc$ before?

-3.10.3 Computing Determinants

For any given matrix, there are a variety of ways to compute the determinant, by "expanding" about any row or column. The determinants of the submatrices used in these computations are used so often that they have their own names. The first is the determinant of a submatrix, the second differs only by a sign.

Definition 94 (Minor In a Matrix) Suppose A is an $n \times n$ matrix and A_{ij} is the $(n-1) \times (n-1)$ submatrix formed by removing row i and column j. Then the **minor** for A at location $i\,j$ is the determinant of the submatrix, $M_{A,ij} = \det(A_{ij})$. $\triangle$

Definition 95 (Cofactor In a Matrix) Suppose A is an $n \times n$ matrix and A_{ij} is the $(n-1) \times (n-1)$ submatrix formed by removing row i and column j. Then the **cofactor** for A at location $i\,j$ is the signed determinant of the submatrix, $C_{A,ij} = (-1)^{i+j} \det(A_{ij})$. $\triangle$

Example 96 (Minors and cofactors) For the matrix,

$$A = \begin{bmatrix} 2 & 4 & 2 & 1 \\ -1 & 2 & 3 & -1 \\ 3 & 1 & 0 & 5 \\ 3 & 6 & 3 & 2 \end{bmatrix}$$

we have minors

$$M_{A,42} = \begin{vmatrix} 2 & 2 & 1 \\ -1 & 3 & -1 \\ 3 & 0 & 5 \end{vmatrix} = 2(15) - 2(-2) + 1(-9) = 25$$
$$M_{A,34} = \begin{vmatrix} 2 & 4 & 2 \\ -1 & 2 & 3 \\ 3 & 6 & 3 \end{vmatrix} = 2(-12) - 4(-12) + 2(-12) = 0$$

and so two cofactors are

$$C_{A,42} = (-1)^{4+2} M_{A,42} = (1)(25) = 25$$
$$C_{A,34} = (-1)^{3+4} M_{A,34} = (-1)(0) = 0$$

A third cofactor is

$$C_{A,12} = (-1)^{1+2} M_{A,12} = (-1) \begin{vmatrix} -1 & 3 & -1 \\ 3 & 0 & 5 \\ 3 & 3 & 2 \end{vmatrix}$$
$$= (-1)((-1)(-15) - (3)(-9) + (-1)(9)) = -33 \qquad \boxtimes$$

With this notation in hand, we can state

Theorem 97 (Determinant Expansion about Rows and Columns) Suppose that A is a square matrix of size n. Then

$$\det(A) = A_{i1}C_{A,i1} + A_{i2}C_{A,i2} + A_{i3}C_{A,i3} + \cdots + A_{in}C_{A,in} \quad 1 \leq i \leq n$$

which is known as **expansion** about row i, and

$$\det(A) = A_{1j}C_{A,1j} + A_{2j}C_{A,2j} + A_{3j}C_{A,3j} + \cdots + A_{nj}C_{A,nj} \quad 1 \leq j \leq n$$

which is known as **expansion** about column j. □

That the determinant of an $n \times n$ matrix can be computed in $2n$ different (albeit similar) ways is nothing short of remarkable. For the doubters among us, we will do an example, computing a 4×4 matrix in two different ways.

Example 98 (Two computations, same determinant) Let

$$A = \begin{bmatrix} -2 & 3 & 0 & 1 \\ 9 & -2 & 0 & 1 \\ 1 & 3 & -2 & -1 \\ 4 & 1 & 2 & 6 \end{bmatrix}$$

Then expanding about the fourth row ($i = 4$) yields,

$$\begin{aligned} |A| &= (4)(-1)^{4+1}\begin{vmatrix} 3 & 0 & 1 \\ -2 & 0 & 1 \\ 3 & -2 & -1 \end{vmatrix} + (1)(-1)^{4+2}\begin{vmatrix} -2 & 0 & 1 \\ 9 & 0 & 1 \\ 1 & -2 & -1 \end{vmatrix} \\ &\quad + (2)(-1)^{4+3}\begin{vmatrix} -2 & 3 & 1 \\ 9 & -2 & 1 \\ 1 & 3 & -1 \end{vmatrix} + (6)(-1)^{4+4}\begin{vmatrix} -2 & 3 & 0 \\ 9 & -2 & 0 \\ 1 & 3 & -2 \end{vmatrix} \\ &= (-4)(10) + (1)(-22) + (-2)(61) + 6(46) = 92 \end{aligned}$$

while expanding about column 3 ($j = 3$) gives

$$\begin{aligned} |A| &= (0)(-1)^{1+3}\begin{vmatrix} 9 & -2 & 1 \\ 1 & 3 & -1 \\ 4 & 1 & 6 \end{vmatrix} + (0)(-1)^{2+3}\begin{vmatrix} -2 & 3 & 1 \\ 1 & 3 & -1 \\ 4 & 1 & 6 \end{vmatrix} + \\ &\quad (-2)(-1)^{3+3}\begin{vmatrix} -2 & 3 & 1 \\ 9 & -2 & 1 \\ 4 & 1 & 6 \end{vmatrix} + (2)(-1)^{4+3}\begin{vmatrix} -2 & 3 & 1 \\ 9 & -2 & 1 \\ 1 & 3 & -1 \end{vmatrix} \\ &= 0 + 0 + (-2)(-107) + (-2)(61) = 92 \end{aligned}$$

⊠

Notice how much easier the second computation was. By choosing to expand about the third column, we have two entries that are zero, so two 3×3 determinants need not be computed at all!

-3.10.4 Properties of Determinants

The determinant is of some interest by itself, but it is of the greatest use when employed to *determine* properties of matrices. To that end, we list some theorems here. Unfortunately, mostly without proof at the moment.

Theorem 99 (Determinant of the Transpose) Suppose that A is a square matrix. Then $\det(A^t) = \det(A)$. □

Proof We will prove this result by induction on the size of the matrix. For a matrix of size 1, the transpose and the matrix itself are equal, so no matter what the definition of a determinant might be, their determinants are equal.

Now suppose the theorem is true for matrices of size $n-1$. By Determinant Expansion about Rows and Columns Theorem we can write the determinant as a product of entries from the first row with their cofactors and then sum these products. These cofactors are signed determinants of matrices of size $n-1$, which by the induction hypothesis, are equal to the determinant of their transposes, and commutativity in the sum in the exponent of -1 means the cofactor is equal to a cofactor of the transpose.

$$\begin{aligned}
\det(A) &= A_{11}C_{A,11} + A_{12}C_{A,12} \\
&\quad + A_{13}C_{A,13} + \cdots + A_{1n}C_{A,1n} && \text{row 1} \\
&= {A^t}_{11}C_{A,11} + {A^t}_{21}C_{A,12} \\
&\quad + {A^t}_{31}C_{A,13} + \cdots + {A^t}_{n1}C_{A,1n} \\
&= {A^t}_{11}C_{A^t,11} + {A^t}_{21}C_{A^t,21} \\
&\quad + {A^t}_{31}C_{A^t,31} + \cdots + {A^t}_{n1}C_{A^t,n1} && \text{Induction hypothesis} \\
&= \det\left(A^t\right) && \text{column 1} \quad \blacksquare
\end{aligned}$$

Theorem 100 (Determinant Respects Matrix Multiplication) Suppose that A and B are square matrices of the same size. Then $\det(AB) = \det(A)\det(B)$. □

It's an amazing thing that matrix multiplication and the determinant interact this way. Might it also be true that $\det(A+B) = \det(A) + \det(B)$?

Theorem 101 (Determinant Respects Matrix Scalar Multiplication) Suppose that $A : n \times n$ is a square matrix and c is a scalar. Then $\det(cA) = c^n \det(A)$. □

Theorem 102 (Determinant of a partitioned matrix) Suppose that $A : n \times n$ is a partitioned matrix and non singular then

$$\begin{aligned} |A| &= \begin{vmatrix} E & F \\ G & H \end{vmatrix} \\ &= |E| \left| H - GE^{-1}F \right| \\ &= |H| \left| E - FH^{-1}G \right| \end{aligned}$$ □

Proof Define the following matrices

$$W = \begin{bmatrix} I & -FH^{-1} \\ \mathcal{O} & I \end{bmatrix} \qquad Z = \begin{bmatrix} I & \mathcal{O} \\ -H^{-1}G & I \end{bmatrix}$$

where $\det(W) = \det(Z) = 1$ since they are triangular. Then we have

$$WAZ = \begin{bmatrix} E - FH^{-1}G & \mathcal{O} \\ \mathcal{O} & H \end{bmatrix}$$

and using Determinant Respects Matrix Multiplication Theorem

$$|A| = |WAZ| = |H| \left| E - FH^{-1}G \right|$$ ■

Theorem 103 (Singular Matrices have Zero Determinants) Let A be a square matrix. Then A is singular if and only if $\det(A) = 0$. □

Theorem 104 (NonSingular Matrix Equivalences, Round 7) Suppose that A is a square matrix of size n. The following are equivalent.

1. A is nonsingular.
2. A row-reduces to the identity matrix.
3. The null space of A contains only the zero vector, $\mathcal{N}(A) = \{\mathbf{0}\}$.
4. The linear system $\mathcal{LS}(A, \boldsymbol{b})$ has a unique solution for every possible choice of $\boldsymbol{b}$.
5. The columns of A are a linearly independent set.
6. A is invertible.
7. The column space of A is $\mathbb{C}^n$, $\mathcal{C}(A) = \mathbb{C}^n$.
8. The columns of A are a basis for $\mathbb{C}^n$.

9. The rank of A is n, $r(A) = n$.

10. The nullity of A is zero, $n(A) = 0$.

11. The determinant of A is nonzero, $\det(A) \neq 0$. □

Proof Singular Matrices have Zero Determinants Theorem says A is singular if and only if $\det(A) = 0$. If we negate each of these statements, we arrive at two contrapositives that we can combine as the equivalence, A is nonsingular if and only if $\det(A) \neq 0$. This allows us to add a new statement to the list. ■

Computationally, row-reducing a matrix is the most efficient way to determine if a matrix is nonsingular, though the effect of using division in a computer can lead to round-off errors that confuse small quantities with critical zero quantities. Conceptually, the determinant may seem the most efficient way to determine if a matrix is nonsingular. The definition of a determinant uses just addition, subtraction and multiplication, so division is never a problem. And the final test is easy: is the determinant zero or not? However, the number of operations involved in computing a determinant very quickly becomes so excessive as to be impractical.

-3.11 Dimension

Definition 105 (Dimension) Suppose that V is a vector space and $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \ldots, \boldsymbol{v}_t\}$ is a basis of V. Then the **dimension** of V is defined by $\dim(V) = t$. If V has no finite bases, we say V has infinite dimension. △

Theorem 106 (Spanning Sets and Linear Dependence) Suppose that $S = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \ldots, \boldsymbol{v}_t\}$ is a finite set of vectors which spans the vector space V. Then any set of $t + 1$ or more vectors from V is linearly dependent. □

Theorem 107 (Bases have Identical Sizes) Suppose that V is a vector space with a finite basis B and a second basis C. Then B and C have the same size. □

Definition 108 (Nullity Of a Matrix) Suppose that A is an $m \times n$ matrix. Then the **nullity** of A is the dimension of the null space of A, $n(A) = \dim(\mathcal{N}(A))$. △

-3.12 Rank of a Matrix

Definition 109 (Rank Of a Matrix) Suppose that A is an $m \times n$ matrix. Then the **rank** of A is the dimension of the column space of A, $r(A) = \dim(\mathcal{C}(A))$. $\triangle$

Theorem 110 (Rules for Ranks) Let $A : m \times n$ be a matrix. Then

1. $0 \leq r(A) \leq \min(m, n)$;
2. $r(A) = r(A^t)$;
3. $r(A + B) \leq r(A) + r(B)$;
4. $r(AB) \leq \min(r(A), r(B))$;
5. $r(AA^t) = r(A^tA) = r(A) = r(A^t)$;
6. For nonsingular $B : m \times m$ and $C : n \times n$, we have $r(BAC) = r(A)$;
7. For $A : n \times n$, $r(A) = n$ if and only if A is nonsingular; $\square$

-3.13 Range and Null Space

Theorem 111
1. $r(A) = \dim(\mathcal{R}(A))$;
2. $\dim(\mathcal{R}(A)) + \dim(\mathcal{N}(A)) = n$;
3. $\mathcal{R}(AA^t) = \mathcal{R}(A)$;
4. $\mathcal{R}(AB) \subseteq \mathcal{R}(A)$ for any A and B; $\square$

-3.14 Definite Matrices and Quadratic Forms

Definition 112 (Quadratic Form) Suppose $A : n \times n$ is symmetric and $\boldsymbol{x}$ is any vector. Then the **quadratic form** in $\boldsymbol{x}$ is defined as the function

$$Q(x) = \boldsymbol{x}^t A \boldsymbol{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \boldsymbol{x}_i \boldsymbol{x}_j \ .$$

Clearly, $Q(\mathbf{0}) = \mathbf{0}$. $\triangle$

Definition 113 (Positive Definite Matrix) The matrix A is called **positive definite** if $Q(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \neq \mathbf{0}$. We write $A > 0$. $\triangle$

If $A > 0$, then $(-A)$ is called **negative definite**.

Theorem 114 Let the $n \times n$ matrix $A > 0$. Then

1. $\boldsymbol{x}^t A \boldsymbol{x} > 0$ for any $\boldsymbol{x} \neq \mathbf{0}$;
2. A is nonsingular and $\det(A) > 0$;
3. $A^{-1} > 0$;
4. $\operatorname{tr} A > 0$;
5. Let $P : n \times m$ be of $r(P) = m \leq n$. Then $P^t A P > 0$ and in particular $P^t P > 0$, choosing $A = I$.
6. Let $P : n \times m$ be of $r(P) < m \leq n$. Then $P^t A P \geq 0$ and in particular $P^t P \geq 0$, choosing $A = I$. □

Theorem 115 (Cauchy–Schwarz Inequality) Let $\boldsymbol{x}$, $\boldsymbol{y}$ be real vectors of same dimension. Then

$$(\langle \boldsymbol{x}, \boldsymbol{y} \rangle)^2 \leq \langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle$$

with equality if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent. □

Theorem 116 Let $\boldsymbol{x}$, $\boldsymbol{y}$ be real vectors and $A > 0$. Then we have the following results:

1. $(\boldsymbol{x}^t A \boldsymbol{y})^2 \leq (\boldsymbol{x}^t A \boldsymbol{x})(\boldsymbol{y}^t A \boldsymbol{y})$;
2. $(\boldsymbol{x}^t \boldsymbol{y})^2 \leq (\boldsymbol{x}^t A \boldsymbol{x})(\boldsymbol{y}^t A^{-1} \boldsymbol{y})$; □

Theorem 117 Let $A > 0$ and T be any square matrix. Then

1. $\sup_{\boldsymbol{x} \neq \mathbf{0}} \frac{(\boldsymbol{x}^t \boldsymbol{y})^2}{\boldsymbol{x}^t A \boldsymbol{x}} = \boldsymbol{y}^t A^{-1} \boldsymbol{y}$;
2. $\sup_{\boldsymbol{x} \neq \mathbf{0}} \frac{(\boldsymbol{y}^t T \boldsymbol{x})^2}{\boldsymbol{x}^t A \boldsymbol{x}} = \boldsymbol{y}^t T A^{-1} T^t \boldsymbol{y}$; □

-3.15 Idempotent Matrices

Definition 118 (Idempotent) A square matrix A is called **idempotent** if it satisfies

$$A^2 = AA = A$$

An idempotent matrix A is called an **orthogonal projector** if $A = A^t$. Otherwise, A is called an **oblique projector** $\triangle$

Theorem 119 Let $A : n \times n$ be idempotent with $r(A) = r \leq n$. Then we have:

1. $\operatorname{tr} A = r(A) = r$;
2. If A is of full rank n, then $A = I_n$;
3. If A and B are idempotent and if $AB = BA$, then AB is also idempotent;
4. If A is idempotent and P is orthogonal, then PAP^t is also idempotent;
5. If A is idempotent, then $I - A$ is idempotent and

$$A(I - A) = (I - A)A = \mathcal{O} \ . \qquad \square$$

-3.16 Projectors

Projectors

Consider the range space $\mathcal{R}(A)$ of the matrix $A : m \times n$ with rank $r = \min m, n$. Then there exists $\mathcal{R}(A)^\perp$, which is the orthogonal complement of $\mathcal{R}(A)$ with dimension $m - r$. Any vector $\boldsymbol{x} \in \mathbb{R}^m$ has the unique decomposition

$$\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2, \qquad \boldsymbol{x}_1 \in \mathcal{R}(A) \ , \text{ and } \boldsymbol{x}_2 \in \mathcal{R}(A)^\perp \ ,$$

of which the component $\boldsymbol{x}_1$ is called the orthogonal projector of $\boldsymbol{x}$ on $\mathcal{R}(A)$. The component $\boldsymbol{x}_1$ can be computed as $P\boldsymbol{x}$, where

$$P = A(A^tA)^{-1}A^t \ ,$$

which is called the **projector operator** on $\mathcal{R}(A)$. Note that P is unique.

Theorem 120 For any $P : n \times n$, the following statements are equivalent:

1. P is an orthogonal projector operator;
2. P is symmetric and idempotent. $\square$

-3.17 Differentiation of Scalar Functions of Matrices

Definition 121 (Partial Differential) If $f(X)$ is a real function of an $m \times n$ matrix X then the partial differential of f with respect to X is defined as the $m \times n$ matrix of partial differentials:

$$\frac{\partial}{\partial X} f(X) = \begin{bmatrix} \frac{\partial}{\partial X_{11}} f & \cdots & \frac{\partial}{\partial X_{1n}} f \\ \vdots & & \vdots \\ \frac{\partial}{\partial X_{m1}} f & \cdots & \frac{\partial}{\partial X_{mn}} f \end{bmatrix} \qquad \triangle$$

Theorem 122 Let $\boldsymbol{x}$ be an n vector and A be a symmetric $n \times n$ matrix. Then

$$\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^t A \boldsymbol{x} = 2A\boldsymbol{x} \ . \qquad \square$$

Proof

$$\boldsymbol{x}^t A \boldsymbol{x} = \sum_{r,s=1}^{n} A_{rs} \boldsymbol{x}_r \boldsymbol{x}_s$$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{x}_i} \boldsymbol{x}^t A \boldsymbol{x} &= \sum_{s=1, s\neq i}^{n} A_{is} \boldsymbol{x}_s + \sum_{r=1, r\neq i}^{n} A_{ri} \boldsymbol{x}_r + 2A_{ii} \boldsymbol{x}_i \\ &= 2 \sum_{s=1}^{n} A_{is} \boldsymbol{x}_s && (\text{as } A_{ij} = A_{ji}) \\ &= 2 \boldsymbol{A}^t{}_i \boldsymbol{x} && (\boldsymbol{A}^t{}_i \text{: } i\text{th row vector of } A) \end{aligned}$$

According to the definition, we get

$$\frac{\partial}{\partial \boldsymbol{x}_i} \boldsymbol{x}^t A \boldsymbol{x} = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{x}_1} \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{x}_n} \end{bmatrix} \boldsymbol{x}^t A \boldsymbol{x} = 2 \begin{bmatrix} \boldsymbol{A}^t{}_1 \\ \vdots \\ \boldsymbol{A}^t{}_n \end{bmatrix} = 2A\boldsymbol{x} \ . \qquad \blacksquare$$

Theorem 123 If $\boldsymbol{x}$ is an n vector, $\boldsymbol{y}$ is an m vector, and C an $n \times m$ matrix, then

$$\frac{\partial}{\partial C} \boldsymbol{x}^t C \boldsymbol{y} = \boldsymbol{x}\boldsymbol{y}^t \ . \qquad \square$$

Proof

$$\boldsymbol{x}^t C \boldsymbol{y} = \sum_{r=1}^{m} \sum_{s=1}^{n} \boldsymbol{x}_s A_{sr} \boldsymbol{y}_r$$

$$\frac{\partial}{\partial C_{kj}} \boldsymbol{x}^t C \boldsymbol{y} = \boldsymbol{x}_k \boldsymbol{y}_j$$

which is the k, jth element of $\boldsymbol{x}\boldsymbol{y}^t$. ■

Theorem 124 If $\boldsymbol{x}$ is an m vector, A a symmetric $n \times n$ and C an $n \times m$ matrix, then

$$\frac{\partial}{\partial C} \boldsymbol{x}^t C^t A C \boldsymbol{x} = 2AC\boldsymbol{x}\boldsymbol{x}^t \ . \qquad \square$$

Proof We have

$$\boldsymbol{x}^t C^t = \left(\sum_{i=1}^{m} \boldsymbol{x}_i C_{1i}, \cdots, \sum_{i=1}^{m} \boldsymbol{x}_i C_{ni} \right)$$

$$\frac{\partial}{\partial C_{kj}} \boldsymbol{x}^t C^t = (0, \cdots, 0, \boldsymbol{x}_j, 0 \cdots, 0) \qquad \text{where } \boldsymbol{x}_j \text{ is in the } k\text{th position.}$$

Using the product rule yields

$$\frac{\partial}{\partial C_{kj}} \boldsymbol{x}^t C^t A C \boldsymbol{x} = \left(\frac{\partial}{\partial C_{kj}} \boldsymbol{x}^t C^t \right) AC\boldsymbol{x} + \boldsymbol{x}^t C^t A \frac{\partial}{\partial C_{kj}} C\boldsymbol{x} \ .$$

Since

$$\boldsymbol{x}^t C^t A = \left(\sum_{j=1}^{n} \sum_{i=1}^{m} \boldsymbol{x}_i C_{ji} A_{j1}, \cdots, \sum_{j=1}^{n} \sum_{i=1}^{m} \boldsymbol{x}_i C_{ji} A_{jn} \right)$$

we get

$$\begin{aligned} \boldsymbol{x}^t C^t A \frac{\partial}{\partial C_{kj}} C\boldsymbol{x} &= \sum_{h,i} \boldsymbol{x}_i \boldsymbol{x}_j C_{hi} A_{kh} \\ &= \sum_{h,i} \boldsymbol{x}_i \boldsymbol{x}_j C_{hi} A_{hk} \qquad \text{as } A \text{ is symmetric} \\ &= \left(\frac{\partial}{\partial C_{kh}} \boldsymbol{x}^t C^t \right) AC\boldsymbol{x} \ . \end{aligned}$$

But $\sum_{h,i} \boldsymbol{x}_i \boldsymbol{x}_j C_{hi} A_{hk}$ is just the k, hth element of the matrix $AC\boldsymbol{x}\boldsymbol{x}^t$. ■

-3.17.1 Differentation of Trace Matrices

Theorem 125 Assume $A = A(x)$ to be a $n \times n$ matrix, where its elements $A_{ij}(x)$ are real functions of a scalar x. Let B be an $n \times n$ matrix, such that its elements are independent of x. Then

$$\frac{\partial}{\partial x} \operatorname{tr} AB = \operatorname{tr} \left(\frac{\partial}{\partial x} A \right) B \ . \qquad \square$$

Proof

$$\operatorname{tr} AB = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} B_{ji}$$

$$\begin{aligned} \frac{\partial}{\partial x} \operatorname{tr} AB &= \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{\partial}{\partial x} A_{ij} \right) B_{ji} \\ &= \operatorname{tr} \left(\frac{\partial}{\partial x} A \right) B \end{aligned} \qquad \blacksquare$$

Theorem 126 For the differentials of the trace we have the following rules:

1. $\partial \operatorname{tr} AX / \partial X = A^t$;
2. $\partial \operatorname{tr} X^t AX / \partial X = (A + A^t) X$;
3. $\partial \operatorname{tr} XAX / \partial X = X^t A + A^t X$;
4. $\partial \operatorname{tr} XAX^t / \partial X = X(A + A^t)$;
5. $\partial \operatorname{tr} X^t AX^t / \partial X = AX^t + X^t A$;
6. $\partial \operatorname{tr} X^t AXB / \partial X = AXB + A^t XB^t$; $\square$

-3.17.2 Differentation of Inverse Matrices

Theorem 127 Let $A = A(x)$ be a regular matrix, such that its elements depend on a scalar x. Then

$$\frac{\partial}{\partial x} A^{-1} = -A^{-1} \frac{\partial}{\partial x} A A^{-1} \qquad \square$$

Proof We have $A^{-1}A = I$, $\partial I/\partial x = \mathcal{O}$, and

$$\frac{\partial}{\partial x}(A^{-1}A) = \left(\frac{\partial}{\partial x}A^{-1}\right)A + A^{-1}\frac{\partial}{\partial x}A = \mathcal{O} \qquad \blacksquare$$

Theorem 128 For nonsingular X, we have

$$\frac{\partial}{\partial X}\operatorname{tr} AX^{-1} = -(X^{-1}AX^{-1})^t$$

$$\frac{\partial}{\partial X}\operatorname{tr} X^{-1}AX^{-1}B = -(X^{-1}AX^{-1}BX^{-1} + X^{-1}BX^{-1}AX^{-1})^t \qquad \square$$

-3.17.3 Differentation of a Determinant

Theorem 129 For a nonsingular matrix X, we have

$$\frac{\partial}{\partial X}|X| = |X|\,(X^t)^{-1}$$

$$\frac{\partial}{\partial X}\log|X| = (X^t)^{-1} \qquad \square$$

Theorem 130 (Kronecker product) Let $A : m \times n$ and $B : p \times q$ be any matrices. Then the **Kronecker product** of A and B is defined as

$$A \otimes B_{ij} = A_{ij}B \qquad\qquad 1 \le i \le m \text{ and } 1 \le j \le n$$

and the following rules hold:

- $c(A \otimes B) = (cA) \otimes B = A \otimes (cB)$;
- $A \otimes (B \otimes C) = (A \otimes B) \otimes C$;
- $A \otimes (B + C) = (A \otimes B) + (A \otimes C)$;
- $(A \otimes B)^t = A^t \otimes B^t$. $\square$

-2 Random Vectors

Definition 131 (Random Vector) Let $Y_1, \cdots, Y_n$ be n random variable defined on the same probability space. The **random vector** $\boldsymbol{Y}$ is

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \triangle$$

Sometime we use the following notation $Y_i = \boldsymbol{Y}_i$.

A Random vector is also called a **multivariate (random) variable**.

Definition 132 (Expected Value) The **expected value** of the random vector $\boldsymbol{Y}$ (if every component exists) is

$$\mathrm{E}(\boldsymbol{Y}) = \begin{bmatrix} \mathrm{E}(Y_1) \\ \mathrm{E}(Y_2) \\ \vdots \\ \mathrm{E}(Y_n) \end{bmatrix} \qquad \triangle$$

Definition 133 (Variance and Covariance Matrix) The **variance and covariance** matrix of the random vector $\boldsymbol{Y}$ (if every component exists) is the $n \times n$ matrix

$$\mathrm{var}(\boldsymbol{Y}) = \begin{bmatrix} \mathrm{var}(Y_1) & \mathrm{cov}(Y_1, Y_2) & \cdots & \mathrm{cov}(Y_1, Y_n) \\ \mathrm{cov}(Y_2, Y_1) & \mathrm{var}(Y_2) & \cdots & \mathrm{cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(Y_n, Y_1) & \mathrm{cov}(Y_n, Y_2) & \cdots & \mathrm{var}(Y_n) \end{bmatrix} \qquad \triangle$$

- $\mathrm{var}(\boldsymbol{Y})$ is symmetric;
- $\mathrm{var}(Y_i) = \mathrm{cov}(Y_i, Y_i)$;
- if the diagonal elements exists then all the other exists (use Cauchy–Schwarz inequality).
- the matrix whose elements are $\mathrm{cor}(Y_i, Y_j) = \mathrm{cov}(Y_i, Y_j)/\sqrt{\mathrm{var}(Y_i)\,\mathrm{var}(Y_j)}$ is called the **correlation matrix**.
- if $\mathrm{var}(\boldsymbol{Y})$ is diagonal then $\boldsymbol{Y}$ is said with **uncorrelated components**.

Let $\boldsymbol{Y}$ be a k vector such that $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $\mathrm{var}(\boldsymbol{Y}) = V$.

Theorem 134 Let $A : n \times k$ a matrix and $\boldsymbol{b}$ a n vector and let

$$\boldsymbol{W} = A\boldsymbol{Y} + \boldsymbol{b} \ .$$

Then $\boldsymbol{W}$ is a n random vector and

1. $\mathrm{E}(\boldsymbol{W}) = A\boldsymbol{\mu} + \boldsymbol{b}$;
2. $\mathrm{var}(\boldsymbol{W}) = AVA^t$. □

Proof Let W_i be the ith component. We prove that $\mathrm{E}(W_i)$ exists.

$$\begin{aligned}
\mathrm{E}(|W_i|) &= \mathrm{E}\left(\left|\sum_{j=1}^{k} A_{ij}Y_j + \boldsymbol{b}_i\right|\right) \\
&\quad \text{linear combination of the } \boldsymbol{Y} \text{ components} \\
&\leq \mathrm{E}\left(\sum_{j=1}^{k} |A_{ij}||Y_j| + |\boldsymbol{b}_i|\right) \\
&\quad \text{by triangle inequality} \\
&= \sum_{j=1}^{k} |A_{ij}|\,\mathrm{E}(|Y_j|) + |\boldsymbol{b}_i| \\
&< \infty \\
&\quad \text{since every components of } \mathrm{E}(|\boldsymbol{Y}|) \text{ is finite}
\end{aligned}$$

Hence

$$\mathrm{E}(W_i) = \mathrm{E}\left(\sum_{j=1}^{k} A_{ij}Y_j + \boldsymbol{b}_i\right)$$

and so

$$\mathrm{E}(\boldsymbol{W}) = A\,\mathrm{E}(\boldsymbol{Y}) + \boldsymbol{b} = A\boldsymbol{\mu} + \boldsymbol{b}$$

For the second statement the proof is quite similar, recall that

$$\mathrm{cov}(A_{ij}X_j, A_{rs}X_s) = A_{ij}A_{rs}\,\mathrm{cov}(X_j, X_s) \ .$$ ■

Theorem 135 The variance matrix V is positive semidefinite and it is positive definite if there no exists a non null vector $\boldsymbol{b}$ such that $\boldsymbol{b}^t\boldsymbol{Y}$ has a degenerate distribution. □

Proof Let $W = \boldsymbol{b}^t\boldsymbol{Y}$; then $0 \leq \text{var}(\boldsymbol{b}^t\boldsymbol{Y}) = \boldsymbol{b}^t V\boldsymbol{b}$. Hence $V \geq 0$. If $\boldsymbol{b}^t V\boldsymbol{b} = 0$ then $\text{var}(\boldsymbol{b}^t\boldsymbol{Y}) = 0$, that is $\boldsymbol{b}^t\boldsymbol{Y}$ is a constant with probability 1. ■

Theorem 136 If $\text{var}(\boldsymbol{Y}) = V > 0$ then there exists a square matrix C of order k such that $\boldsymbol{W} = C\boldsymbol{Y}$ have uncorrelated components with unit variance, that is, $\text{var}(\boldsymbol{W}) = I_k$. □

Proof The matrix V can be decomposed as $V = BB^t$ (could you say why?). We let $\boldsymbol{W} = B^{-1}\boldsymbol{Y}$, then

$$\begin{aligned}\text{var}(\boldsymbol{W}) &= \text{var}(B^{-1}\boldsymbol{Y}) = B^{-1}\,\text{var}(\boldsymbol{Y})(B^{-1})^t \\ &= B^{-1}V(B^t)^{-1} = B^{-1}BB^t(B^t)^{-1} \\ &= I_k\end{aligned}$$

■

Theorem 137 Let A a square matrix of order k. Then

$$\text{E}(\boldsymbol{Y}^t A\boldsymbol{Y}) = \boldsymbol{\mu}^t A\boldsymbol{\mu} + \text{tr}(AV)\ .$$ □

Proof

$$\begin{aligned}\text{E}(\boldsymbol{Y}^t A\boldsymbol{Y}) &= \text{E}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} Y_i A_{ij} Y_j\right) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} A_{ij}\,\text{E}(Y_iY_j) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} A_{ij}(\mu_i\mu_j + V_{ij}) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} A_{ij}\mu_i\mu_j + \sum_{i=1}^{k}\sum_{j=1}^{k}(A_{ij}V_{ji}) \qquad V \text{ is symmetric} \\ &= \boldsymbol{\mu}^t A\boldsymbol{\mu} + \sum_{i=1}^{k}[AV]_{ii} \\ &= \boldsymbol{\mu}^t A\boldsymbol{\mu} + \text{tr}(AV)\end{aligned}$$

An alternative proof is as follows:

$$\begin{aligned}\text{E}(\boldsymbol{Y}^t A\boldsymbol{Y}) &= \text{E}(\text{tr}(\boldsymbol{Y}^t A\boldsymbol{Y})) \\ &= \text{E}(\text{tr}(A\boldsymbol{Y}\boldsymbol{Y}^t)) \\ &= \text{tr}\left(A\,\text{E}(\boldsymbol{Y}\boldsymbol{Y}^t)\right) \\ &= \text{tr}\left(A(\boldsymbol{\mu}\boldsymbol{\mu}^t + V)\right) \\ &= \boldsymbol{\mu}^t A\boldsymbol{\mu} + \text{tr}(AV)\end{aligned}$$

■

-1 Multivariate Normal Distribution

-1.1 Multivariate Normal Distribution

Theorem 138 (Linear Combination of Normal Random Variables) Let $Y_i \sim N(\mu_i, \sigma_i^2)$ be k normal random variables such that Y_i is independent from Y_j $(j \neq i)$ and let a_i and b_i constants, then

$$\sum_{i=1}^{n}(a_i Y_i + b_i) \sim N\left(\sum_{i=1}^{k}(a_i\mu_i + b_i), \sum_{i=1}^{k} a_i^2\sigma_i^2\right) \qquad \square$$

In a more compact form we can write (with obvious notation)

$$\langle \boldsymbol{a},\, \boldsymbol{Y}\rangle + \langle \mathbf{1},\, \boldsymbol{b}\rangle \sim N(\langle \boldsymbol{a},\, \boldsymbol{\mu}\rangle + \langle \mathbf{1},\, \boldsymbol{b}\rangle\,, \boldsymbol{a}^t V \boldsymbol{a})$$

where $V = \operatorname{diag}(\sigma_1^2, \cdots, \sigma_k^2)$.

Definition 139 (Multivariate Normal Standard Distribution) Let $Z_i \sim N(0,1)$ be k independent normal standard random variables. The random vector $\boldsymbol{Z}$ with density:

$$\begin{aligned} f(\boldsymbol{Z} = \boldsymbol{z}) &= f(Z_1 = z_1, Z_2 = z_2, \cdots, Z_k = z_k) \\ &= \prod_{i=1}^{k} f(Z_i = z_i) \\ &= (2\pi)^{-k/2} \exp\left(-\frac{1}{2}\boldsymbol{z}^t\boldsymbol{z}\right) \end{aligned}$$

(since the stochastic independence of the components) is called the **Multivariate Normal Standard** variable. $\triangle$

Definition 140 (Multivariate Normal Distribution) Let $\boldsymbol{Z}$ be a multivariate normal standard distribution, A be a $k \times k$ matrix and $\boldsymbol{\mu}$ a k vector. Then

$$\boldsymbol{Y} = A\boldsymbol{Z} + \boldsymbol{\mu}$$

is a **multivariate normal** variable. $\triangle$

What is the density of $\boldsymbol{Y}$?

Since

$$\boldsymbol{Z} = A^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$$

then the Jacobian of the transformation is

$$\left|\frac{\partial}{\partial Y_j} Z_i\right| = |A|^{-1} = |V|^{-1/2}$$

where

$$|V| = \left|AA^t\right| = |A|^2$$

Then, let $\boldsymbol{y} = A\boldsymbol{z} + \boldsymbol{\mu}$ and

$$\begin{aligned} \boldsymbol{z}^t\boldsymbol{z} &= \left(A^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)^t \left(A^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) \\ &= (\boldsymbol{y}-\boldsymbol{\mu})^t V^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \end{aligned}$$

hence the density of $\boldsymbol{Y}$ at $\boldsymbol{y}$ is

$$f(\boldsymbol{Y}=\boldsymbol{y}) = (2\pi)^{-k/2}\,|V|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^t V^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right).$$

The contour line of the density is such that $(\boldsymbol{y}-\boldsymbol{\mu})^t V^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) = c$ which is the equation of an ellipsoid.

Theorem 141 Given $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, V)$. If the matrix V is diagonal then the r.v. $Y_1, Y_2, \cdots, Y_k$ are independent. □

Note that the two facts:

1. $Y_1, Y_2, \cdots, Y_k$ are normal distributed;
2. $Y_1, Y_2, \cdots, Y_k$ are uncorrelated,

does not imply that $\boldsymbol{Y}^t = (Y_1, Y_2, \cdots, Y_k)$ is a multivariate normal random vector as shown by the next example.

Example 142 Let $Z \sim N(0,1)$ and

$$Y = \begin{cases} -Z & \text{if } |Z| < c \\ Z & \text{if } |Z| \geq c \end{cases}$$

for a positive constant c. Since the particular type of the transformation the conditional distribution of $Y|Z=z$ is degenerate (and hence, Z and Y are not independent). The marginal distribution of Y is

$$f(Y=z) = \begin{cases} f(Z=-z) & \text{if } |z| < c \\ f(Z=z) & \text{if } |z| \geq c \end{cases}$$

and since the symmetry of f around 0 we have that $f(Y = z) = f(Z = z)$ regardless the value of c, that is, $Y \sim N(0, 1)$. But the join density of (Y, Z) is

$$\begin{aligned} f(Y = y, Z = z) = f(z)\,\mathrm{I}(y = z)\,\mathrm{I}([|z| \geq c]) \\ + f(z)\,\mathrm{I}(y = -z)\,\mathrm{I}(|z| < c) \end{aligned}$$

which is not the product of two normal density and hence the variables are dependent. There exists a c such that $\mathrm{cov}(Y, Z) = 0$ in fact

$$\begin{aligned} \mathrm{cov}(Y, Z) &= \mathrm{E}(YZ) \\ &= 2\int_c^{+\infty} z^2 f(z)dz - 2\int_0^c z^2 f(z)dz \\ &= 1 - 4\int_0^c z^2 f(z)dz \end{aligned}$$

which is a continuous function of c. For $c = 0$ the $\mathrm{cov}(Y, Z) > 0$ while for $c \to \infty$ the $\mathrm{cov}(Y, Z) < 0$ and hence there must exists a c such that $\mathrm{cov}(Y, Z) = 0$. In particular c is such that

$$\int_c^{+\infty} z^2 f(z)dz = \int_0^c z^2 f(z)dz$$

and by a change of variable formula ($u = z^2$) is

$$\int_c^{+\infty} \frac{1}{2\sqrt{2\pi}} u^{\frac{1}{2}} e^{-\frac{1}{2}u} du = \int_0^c \frac{1}{2\sqrt{2\pi}} u^{\frac{1}{2}} e^{-\frac{1}{2}u} du$$

and hence c is the median of a χ^2_3 distribution, i.e. $c = 2$ (Note: we have to adjust by a suitable constant, both side: $\sqrt{\pi}/\Gamma(3/2)$). ⊠

Theorem 143 (Moments of the Multivariate Normal variable)

$$\mathrm{E}(\boldsymbol{Y}) = \mathrm{E}(A\boldsymbol{Z} + \boldsymbol{\mu}) = \boldsymbol{\mu}$$

$$\mathrm{var}(\boldsymbol{Y}) = \mathrm{var}(A\boldsymbol{Z} + \boldsymbol{\mu}) = V$$

since $\mathrm{E}(\boldsymbol{Z}) = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{Z}) = I_k$. □

Since this theorem we can use the following notation $\boldsymbol{Y} \sim N_k(\boldsymbol{\mu}, V)$.

Theorem 144 Let $\boldsymbol{W} = B\boldsymbol{Y} + \boldsymbol{b}$ and $\boldsymbol{Y} = A\boldsymbol{Z} + \boldsymbol{\mu}$ then

$$\boldsymbol{W} = BA\boldsymbol{Z} + (B\boldsymbol{\mu} + \boldsymbol{b}), \qquad (BA)(BA)^t = BVB^t$$

and

$$W \sim N_k(B\boldsymbol{\mu} + b, BVB^t)\ . \qquad \square$$

Example 145 (Bivariate Normal Variables) The contour line for $k = 2$ is an ellipse, and the angle between the principal axis with the axis of the first component is

$$\omega = \frac{1}{2} \arctan\left(\frac{2V_{12}}{V_{11} - V_{22}}\right)$$

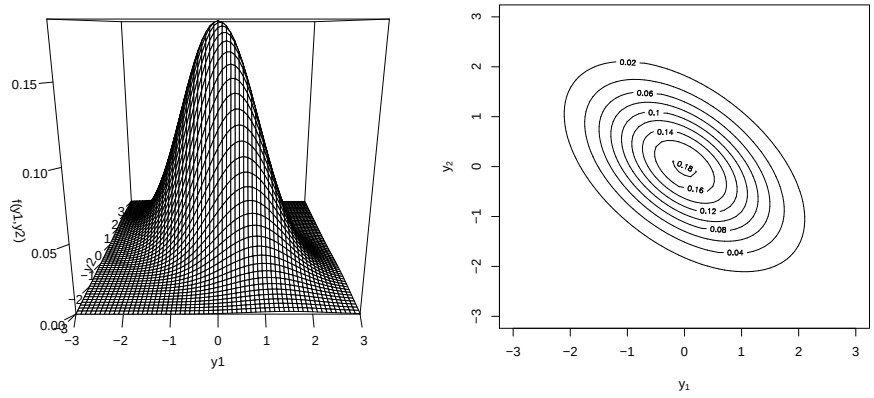
Let

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad V = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} ,$$

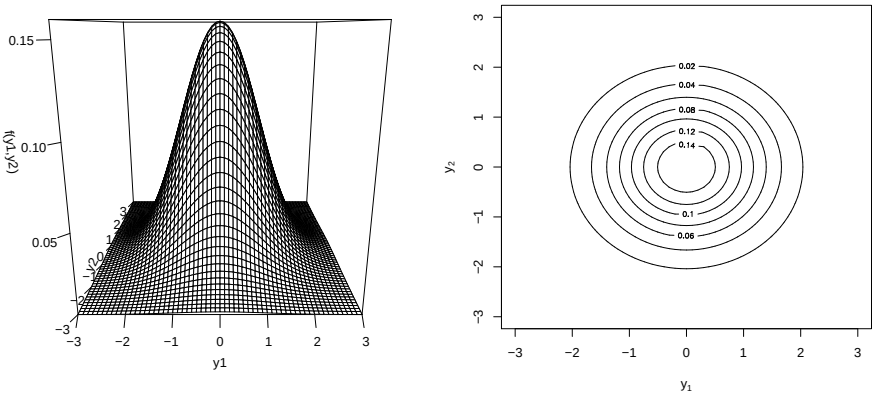
then the density is

$$f(y_1, y_2) = 2\pi(1-\rho^2)^{-1/2} \exp\left(-\frac{1}{2(1-\rho^2)}(y_1^2 - 2\rho y_1 y_2 + y_2^2)\right) .$$

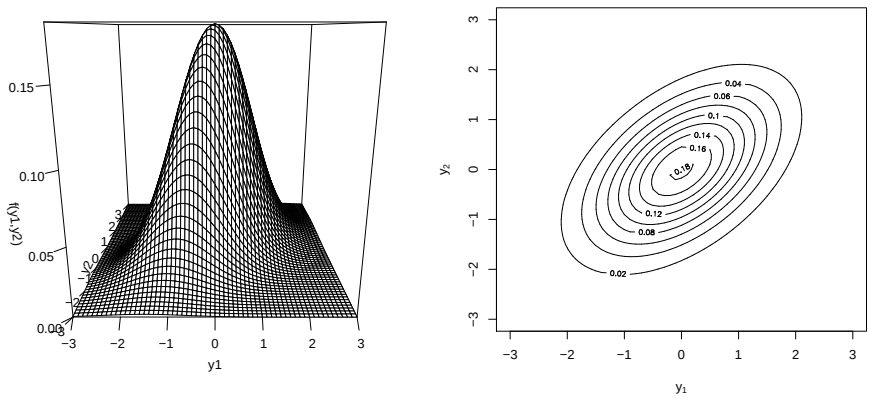
In this case $\omega = \text{sign}(\rho)\pi/4$.



The bivariate normal density for $\rho = -0.5$, in a 3D plot (left) and by contour lines (right)



The bivariate normal density for $\rho = 0$, in a 3D plot (left) and by contour lines (right)



The bivariate normal density for $\rho = +0.5$, in a 3D plot (left) and by contour lines (right) ⊠

Theorem 146 (Marginal distributions) Let $\boldsymbol{Y}$ be a k multivariate normal random vector. And

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix}$$

with $\boldsymbol{Y}_1 \in \mathbb{R}^r$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{k-r}$ $(0 < r < k)$. In the same way

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad , \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \quad V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

with $V_{21} = {V_{21}}^t$. The marginal distribution of $\boldsymbol{Y}_1$ is

$$\boldsymbol{Y}_1 \sim N_r(\boldsymbol{\mu}_1, V_{11}) \ . \qquad \square$$

The vice versa may not holds.

Theorem 147 (Conditional distributions) Let $\boldsymbol{Y}$ be a k multivariate normal random vector. And

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix}$$

with $\boldsymbol{Y}_1 \in \mathbb{R}^r$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{k-r}$ $(0 < r < k)$. In the same way

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad , \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \quad V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

with $V_{21} = {V_{21}}^t$. The conditional distribution of $\boldsymbol{Y}_1|\boldsymbol{Y}_2 = \boldsymbol{y}_2$ is

$$\boldsymbol{Y}_1|\boldsymbol{Y}_2 = \boldsymbol{y}_2 \sim N_r(\boldsymbol{\mu}_1 + V_{12}V_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), V_{11\cdot 2}) \ .$$

where $V_{11\cdot 2} = V_{11} - V_{12}V_{22}^{-1}V_{21}$. □

Proof We first let

$$V^{-1} = W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$$

so that $W_{11} = V_{11} - V_{12}V_{22}^{-1}V_{21}^{-1}$, $W_{22} = V_{22} - V_{21}V_{11}^{-1}V_{12}^{-1}$, $W_{12} = -W_{11}V_{12}V_{22}^{-1}$ and $W_{21} = -W_{22}V_{21}V_{11}^{-1}$. The conditional density of $\boldsymbol{Y}_1|\boldsymbol{Y}_2 = \boldsymbol{y}_2$ is proportional to

$$\exp\left(-\frac{1}{2}Q\right) = \exp\left(-\frac{1}{2}\left(\boldsymbol{z}^t V^{-1}\boldsymbol{z} - {\boldsymbol{z}_2}^t V_{22}^{-1}\boldsymbol{z}_2\right)\right)$$

where we let $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{\mu}$ and $\boldsymbol{z}_2 = \boldsymbol{y}_2 - \boldsymbol{\mu}_2$. Then

$$\begin{aligned} Q &= ({\boldsymbol{z}_1}^t, {\boldsymbol{z}_2}^t) \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{bmatrix} - {\boldsymbol{z}_2}^t \left(W_{22} - W_{21}W_{11}^{-1}W_{12}\right)\boldsymbol{z}_2 \\ &= \left(\boldsymbol{z}_1 + W_{11}^{-1}W_{12}\boldsymbol{z}_2\right)^t W_{11}\left(\boldsymbol{z}_1 + W_{11}^{-1}W_{12}\boldsymbol{z}_2\right) \\ &= \left(\boldsymbol{z}_1 - V_{12}V_{22}^{-1}\boldsymbol{z}_2\right)^t V_{11\cdot 2}^{-1}\left(\boldsymbol{z}_1 - V_{12}V_{22}^{-1}\boldsymbol{z}_2\right) \end{aligned}$$

which is a quadratic form in $\boldsymbol{y}_1$ for each fixed $\boldsymbol{y}_2$. ■

-1.2 Mahalanobis Distance

Definition 148 (Mahalanobis Distance) Given a variance matrix V. The metrics induced by the norm

$$\|\boldsymbol{a}\|_V^2 = \boldsymbol{a}^t V^{-1}\boldsymbol{a}$$

for $\boldsymbol{a} \in \mathbb{R}^k$ is called the **Mahalanobis distance** of the vector $\boldsymbol{a}$. △

1. When $V = I$ then it is the usual Euclidean distance;
2. it is the distance of $\boldsymbol{a}$ from the origin of the axes;
3. if C is a matrix, let $\boldsymbol{b} = C\boldsymbol{a}$ and $W = CVC^t$ then a change in the coordinates by C leads to

   $$\begin{aligned} \|\boldsymbol{b}\|_W^2 &= \boldsymbol{b}^t W^{-1}\boldsymbol{b} = \boldsymbol{a}^t C^t (C^t)^{-1} V^{-1} C^{-1} C\boldsymbol{a} \\ &= \boldsymbol{a}^t V^{-1}\boldsymbol{a} = \|\boldsymbol{a}\|_V^2 \end{aligned}$$

   that is the Mahalanobis distance of the vector is independent of the coordinate system in use;

4. there is a strong relation with the density of a multivariate normal random vector, in fact, the density of $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, V)$ could be written as

$$f(\boldsymbol{Y} = \boldsymbol{y}) = (2\pi)^{-k/2} |V|^{-1/2} \exp\left(-\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\mu}\|_V^2\right)$$

-1.3 Noncentral Chi-square

Definition 149 (Chi-square) Let $\boldsymbol{Z} \sim N(\boldsymbol{0}, I)$ then

$$U_k = \langle \boldsymbol{Z}, \boldsymbol{Z} \rangle = \sum_{i=1}^{k} Z_i^2$$

is a **chi-square** random variable (χ_k^2) with k degree of freedom. △

Definition 150 (Noncentral Chi-square) Let $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, I)$ then

$$U_k = \langle \boldsymbol{Z}, \boldsymbol{Z} \rangle = \sum_{i=1}^{k} Z_i^2$$

is a **noncentral chi-square** random variable ($\chi_k^2(\lambda)$) with k degree of freedom and non centrality parameter $\lambda = \frac{1}{2} \langle \boldsymbol{\mu}, \boldsymbol{\mu} \rangle$. △

We have

$$\begin{aligned} \mathrm{E}(U_k) &= k + 2\lambda \\ \mathrm{var}(U_k) &= 2\ k + 8\ lambda \end{aligned}$$

Theorem 151 (Sum of Noncentral Chi-square) If $W_1, W_2, \cdots W_k$ are jointly independent and if each is distributed as the non central chi-square, so that W_i has p_i degrees of freedom and non centrality parameter λ_i, then $W = \sum_{i=1}^{k} W_i$ has the non central chi-square distribution with $p = \sum_{i=1}^{k} p_i$ degrees of freedom and $\lambda = \sum_{i=1}^{k} \lambda_i$ non centrality parameter. □

Theorem 152 If a vector $\boldsymbol{Y}$ is distributed $N(\boldsymbol{\mu}, \sigma^2 I)$ then $\langle \boldsymbol{Y}, \boldsymbol{Y} \rangle / \sigma^2$ has the non central chi-square distribution with $\lambda = \langle \boldsymbol{\mu}, \boldsymbol{\mu} \rangle / (2\sigma^2)$. □

Theorem 153 If a vector $\boldsymbol{Y}$ is distributed $N(\boldsymbol{\mu}, D)$ where D is diagonal, then $\boldsymbol{Y}^t D^{-1} \boldsymbol{Y}$ has the non central chi-square distribution with k degrees of freedom and parameter $\lambda = \frac{1}{2} \boldsymbol{\mu}^t D^{-1} \boldsymbol{\mu}$. □

-1.4 Distribution of Quadratic Forms

Theorem 154 If the random vector $\boldsymbol{Y}$ is distributed $N(\mathbf{0}, I)$, a necessary and sufficient condition that the quadratic form $\boldsymbol{Y}^t A \boldsymbol{Y}$ be distributed as χ_k^2 is that A be an idempotent matrix of rank k. □

Proof We proof only sufficiency. If A is idempotent of rank k, there exists an orthogonal matrix P such that $P^t A P = \begin{bmatrix} I_k & \mathcal{O} \\ \mathcal{O} & \mathcal{O} \end{bmatrix}$. Define $\boldsymbol{Z} = P^t Y$. Then $\boldsymbol{Y}^t A \boldsymbol{Y} = Z^t P^t A P Z = \boldsymbol{Z}_1{}^t \boldsymbol{Z}_1$, where $\boldsymbol{Z}^t = (\boldsymbol{Z}_1{}^t, \boldsymbol{Z}_2{}^t)$. Hence $\boldsymbol{Z} = N(\mathbf{0}, I)$, and it follows that $\boldsymbol{Z}_1{}^t \boldsymbol{Z}_1 = \sum_{i=1}^k \boldsymbol{Z}_{1i}{}^2$, which is a sum of squares of independent normal variables with means 0 and variances 1, is distributed as χ_k^2. ■

Theorem 155 If $\boldsymbol{Y}$ is distributed $N(\boldsymbol{\mu}, I)$, then $\boldsymbol{Y}^t A \boldsymbol{Y}$ is distributed as $\chi_k^2(\lambda)$, where $\lambda = 1/2 \boldsymbol{\mu}^t A \boldsymbol{\mu}$, if and only if A is an idempotent matrix of rank k. □

Theorem 156 If a vector $\boldsymbol{Y}$ is distributed $N(\mathbf{0}, V)$ then $Y^t B Y$ is distributed as χ_k^2 if and only if BV is idempotent of rank k. □

Theorem 157 If a vector $\boldsymbol{Y}$ is distributed $N(\boldsymbol{\mu}, V)$ then $Y^t B Y$ is distributed as $\chi_k^2(\lambda)$, where $\lambda = \frac{1}{2} \boldsymbol{\mu}^t B \boldsymbol{\mu}$ and k is the rank of B if and only if BV is idempotent. □

-1.5 Independence of Quadratic Forms

Theorem 158 If $\boldsymbol{Y}$ is distributed $N(\boldsymbol{\mu}, \Sigma)$, the two positive semidefinite quadratic forms $\boldsymbol{Y}^t A \boldsymbol{Y}$ and $\boldsymbol{Y}^t B \boldsymbol{Y}$ are independent if and only if $A \Sigma B = \mathcal{O}$. □

Theorem 159 If $\boldsymbol{Y}$ is distributed $N(\boldsymbol{\mu}, I)$, the p positive semidefinite quadratic forms $\boldsymbol{Y}^t B_1 \boldsymbol{Y}$, $\boldsymbol{Y}^t B_2 \boldsymbol{Y}$, $\cdots$, $\boldsymbol{Y}^t B_p \boldsymbol{Y}$ are jointly independent if and only if $B_i B_j = \mathcal{O}$ for all $i \neq j$. □

Theorem 160 Let $A = \sum_{i=1}^p A_i$ be a $k \times k$ matrix and the A_i symmetric. If $\boldsymbol{Y}$ is distributed $N_k(\boldsymbol{\mu}, I)$ and if $\boldsymbol{Y}^t A \boldsymbol{Y} = \sum_{i=1}^p \boldsymbol{Y}^t A_i \boldsymbol{Y}$, a necessary and sufficient condition that $\boldsymbol{Y}^t A_i \boldsymbol{Y}$ be distributed as $\chi_{n_i}^2(\lambda_i)$, where n_i is the rank of A_i and $\lambda_i = \frac{1}{2} \mu^t A_i \mu$, and for the $\boldsymbol{Y}^t A_1 \boldsymbol{Y}$, $\boldsymbol{Y}^t A_2 \boldsymbol{Y}$, $\cdots$, $\boldsymbol{Y}^t A_p \boldsymbol{Y}$ to be jointly independent, is that the rank of A be equal to the sum of the ranks of separate A_i; that is to say, that $\sum_{i=1}^p r(A_i) = r(\sum_{i=1}^p A_i) = r(A)$. □

The famous Cochran–Fisher theorem is a special case of this Theorem when $\boldsymbol{\mu} = \mathbf{0}$.

Further the previous theorem holds if two out of three of the following conditions hold

- A_i are idempotent;
- $A_i A_j = \mathcal{O}$ for all $i \neq j$;
- A is idempotent.

Theorem 161 If $\boldsymbol{Y}$ is distributed $N(\boldsymbol{\mu}, \sigma^2 I)$ the positive semidefinite quadratic forms $\boldsymbol{Y}^t A \boldsymbol{Y}$ and $\boldsymbol{Y}^t B \boldsymbol{Y}$ are independent if $\operatorname{tr} BA = 0$, or in other, words if the covariance of $\boldsymbol{Y}^t A \boldsymbol{Y}$ and $\boldsymbol{Y}^t B \boldsymbol{Y}$ equal zero. □

Theorem 162 (Independence of Linear and Quadratic Forms) If B is a $q \times n$ matrix, A is an $n \times n$ matrix, and $\boldsymbol{Y}$ is distributed $N(\boldsymbol{\mu}, \sigma^2 I)$, then the linear forms $B\boldsymbol{Y}$ are independent of the quadratic form $\boldsymbol{Y}^t A \boldsymbol{Y}$ if $BA = \mathcal{O}$. □

Theorem 163 If $\boldsymbol{Y}$ is distributed with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I$, the expected value of the quadratic form $\boldsymbol{Y}^t A \boldsymbol{Y}$ is equal to $\sigma^2 \operatorname{tr} A$. □

Theorem 164 If $\boldsymbol{Y}$ is distributed with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I$ and A is an idempotent matrix of rank p then the expected value of the quadratic form $\boldsymbol{Y}^t A \boldsymbol{Y}$ is equal to $\sigma^2 p$. □

0 Simple Examples

Picea Abiens* and *Pinus Sylvestris

In the 1999 we had measure the diameters and the heights of 190 trees of the species *Picea Abiens* and *Pinus Sylvestris* in San Vito di Cadore. 136 measures are related to *Picea Abiens* while the remains 54 to *Pinus Sylvestris*.

The diameter is measured using a tools call tresle (cavalletto). And the measure is taken $1.30m$ from the ground. Only trees with diamter bigger than $18cm$ are reported.

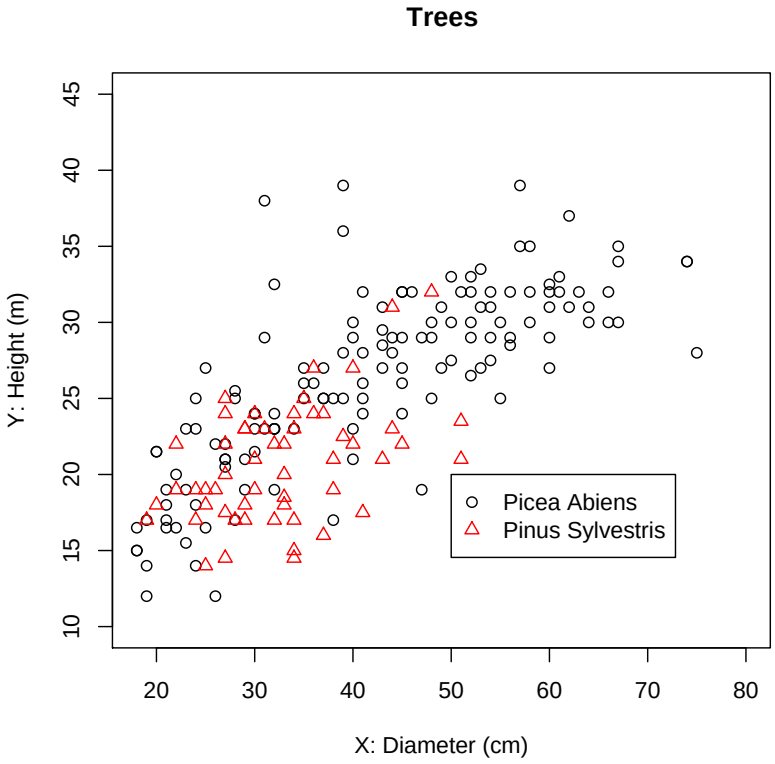
Since the measure of the height involve an instrument with high precision and the operation need some time, we would like to build a model to predict the height of a tree given its diameter.

Load the data into R

```
> abeti <- read.table(file = "../data/abete-rosso.dat",
+     sep = ",", header = FALSE)
> pini <- read.table(file = "../data/pino-silvestre.dat",
+     sep = ",", header = FALSE)
> a.diametro <- abeti[, 1]
> a.altezza <- abeti[, 2]
> p.diametro <- pini[, 1]
> p.altezza <- pini[, 2]
> diametro <- c(a.diametro, p.diametro)
> altezza <- c(a.altezza, p.altezza)
> albero <- c(rep(0, length(a.diametro)),
+     rep(1, length(p.diametro)))
```

Plot the data

```
> plot(a.diametro, a.altezza, xlim = c(18,
+     80), ylim = c(10, 45), pch = 1,
+     col = 1, xlab = "X: Diameter (cm)",
+     ylab = "Y: Height (m)", main = "Trees")
> points(p.diametro, p.altezza, pch = 2,
+     col = 2)
> legend(x = 50, y = 20, legend = c("Picea Abiens",
+     "Pinus Sylvestris"), pch = c(1,
+     2), col = c(1, 2))
```



First model

Let Z be the indicator variable that is equal to 0 when the measure is related to a *Picea Abiens* and 1 otherwise.

$$Y = \alpha + \beta X + \gamma Z + \varepsilon$$

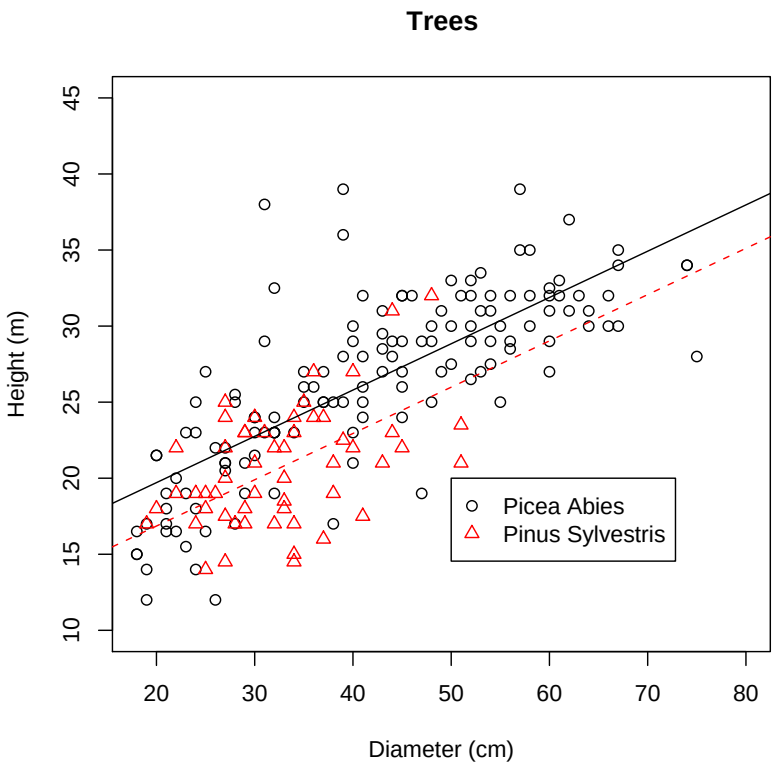
so that, we have

$$\begin{array}{lcl} \textit{Picea Abiens} & : & Y = \alpha + \beta X + \varepsilon \\ \textit{Pinus Sylvestris} & : & Y = (\alpha + \gamma) + \beta X + \varepsilon \end{array}$$

Fit the first model

```
> modello.a <- lm(altezza ~ diametro +
+     albero)
> modello.a

Call:
lm(formula = altezza ~ diametro + albero)
```

```
Coefficients:
(Intercept)     diametro       albero
    13.6221       0.3043      -2.8543
```



Summary statistics for the first model

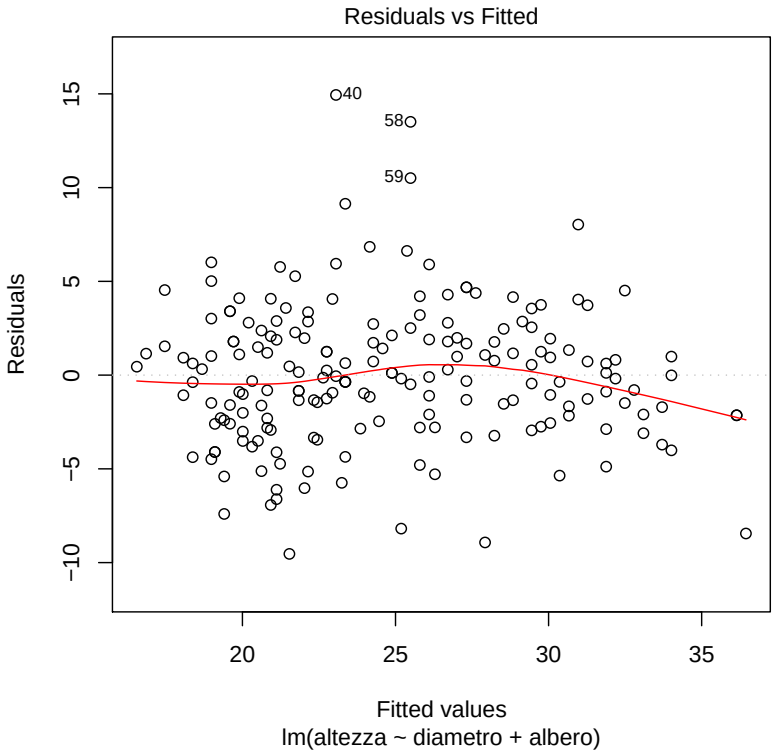
```
> summary(modello.a)

Call:
lm(formula = altezza ~ diametro + albero)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5345 -2.3808 -0.1178  1.9839 14.9439

Coefficients:
            Estimate Std. Error t value
(Intercept) 13.62215    0.93229  14.611
```

```
diametro     0.30432    0.02092  14.548
albero      -2.85434    0.63136  -4.521
            Pr(>|t|)
(Intercept)  < 2e-16 ***
diametro     < 2e-16 ***
albero      1.09e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.742 on 187 degrees of freedom
Multiple R-squared: 0.6152,        Adjusted R-squared: 0.6111
F-statistic: 149.5 on 2 and 187 DF,  p-value: < 2.2e-16
```

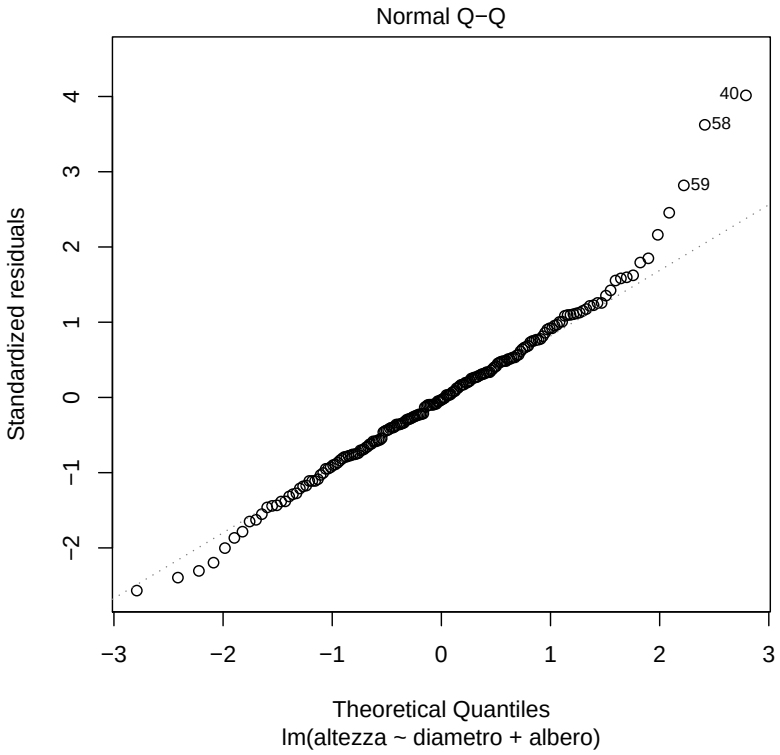
Default plot for the first model

```
> plot(modello.a, which = 1)

> plot(modello.a, which = 2)

> plot(modello.a, which = 3)

> plot(modello.a, which = 5)
```

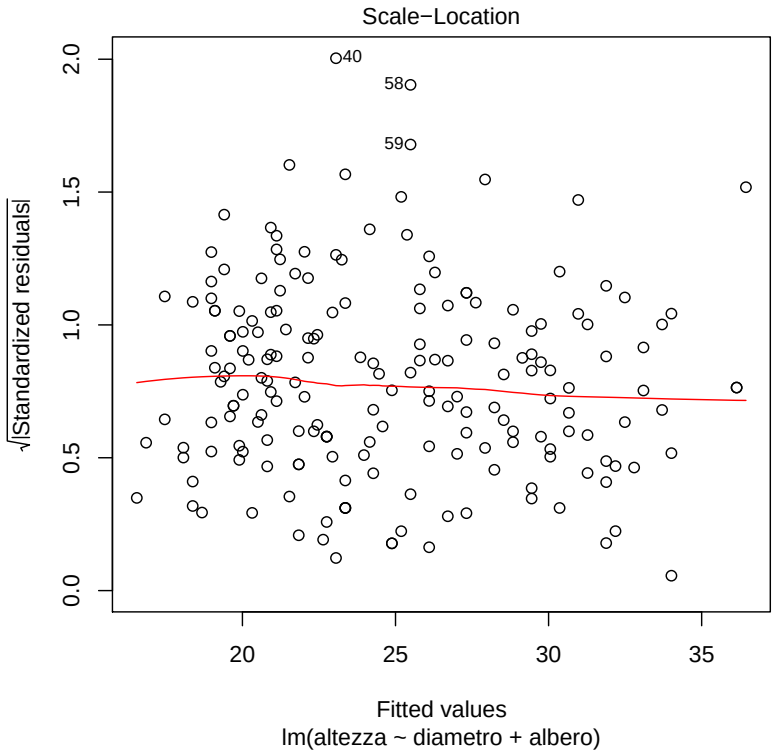
Residuals vs Fitted Values



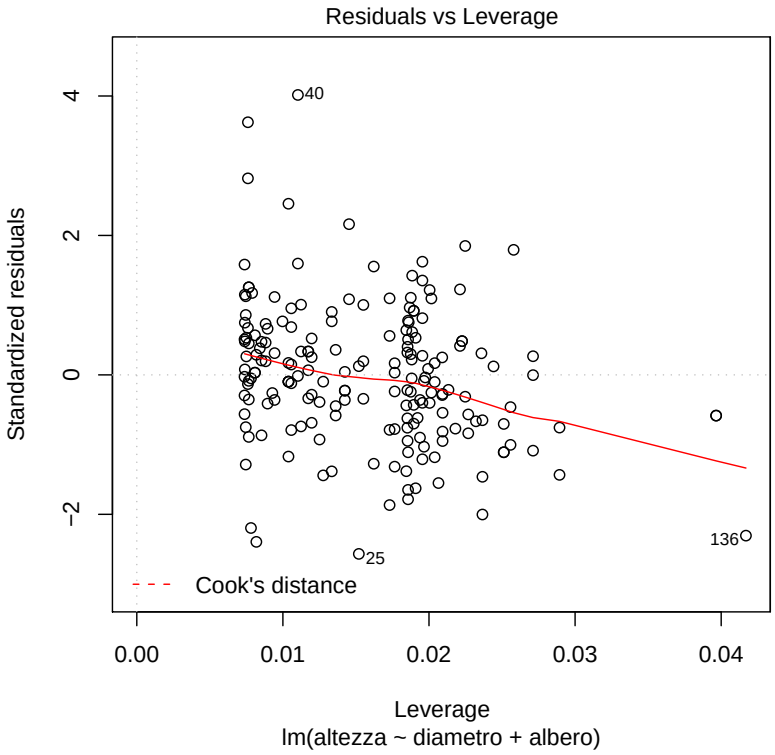
Standardized Residuals vs Theoretical Quantiles



Squared root of Standardized Residuals vs Fitted Values



Standardized Residuals vs Leverage Points



Second model

Let Z be the indicator variables that is equal to 0 when the measure is related to a *Picea Abiens* and 1 otherwise.

$$Y = \alpha\sqrt{X} + \beta(\sqrt{X} * Z) + \varepsilon$$

so that, we have

$$\begin{array}{lcl} \textit{Picea Abiens} & : & Y = \alpha\sqrt{X} + \varepsilon \\ \textit{Pinus Sylvestris} & : & Y = (\alpha + \beta)\sqrt{X} + \varepsilon \end{array}$$

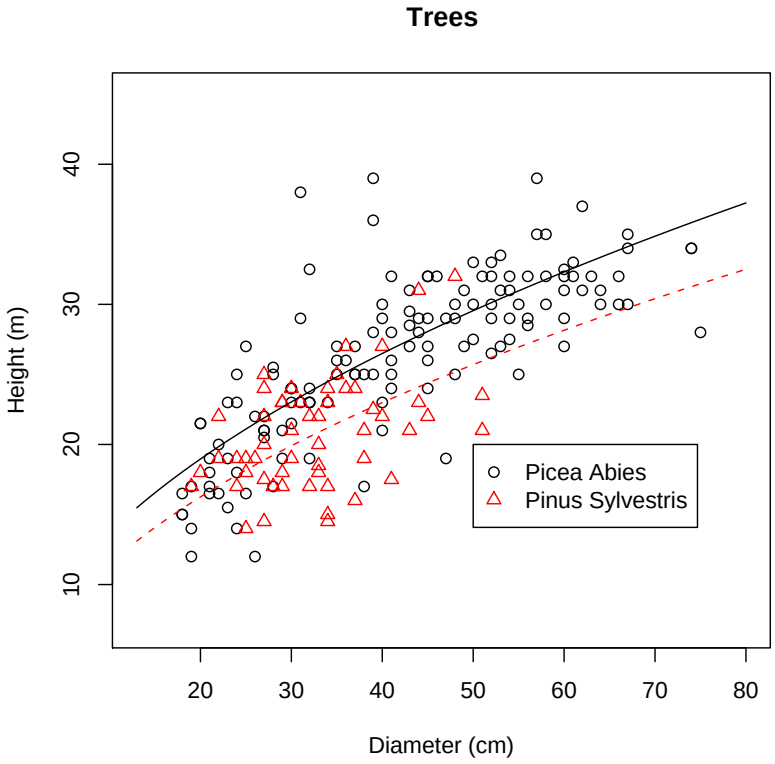
Fit the second model

```
> modello.d <- lm(altezza ~ I(sqrt(diametro)) +
+     I(sqrt(diametro)):albero -
+     1)
> modello.d.abete <- function(x) {
+     return(modello.d$coefficients[1] *
+         sqrt(x + 1))
```

```
+ }
> modello.d.pino <- function(x) {
+     return(sum(modello.d$coefficients) *
+         sqrt(x))
+ }
```

Plot the second model

```
> plot(a.diametro, a.altezza, xlim = (range(diametro) +
+     c(-5, 5)), ylim = c(7, 45),
+     pch = 1, col = 1, xlab = "Diameter (cm)",
+     ylab = "Height (m)", main = "Trees")
> points(p.diametro, p.altezza, pch = 2,
+     col = 2)
> legend(x = 50, y = 20, legend = c("Picea Abies",
+     "Pinus Sylvestris"), pch = c(1,
+     2), col = c(1, 2))
> curve(modello.d.abete, lty = 1,
+     col = 1, add = TRUE)
> curve(modello.d.pino, lty = 2,
+     col = 2, add = TRUE)
```



Summary statistics for the second model

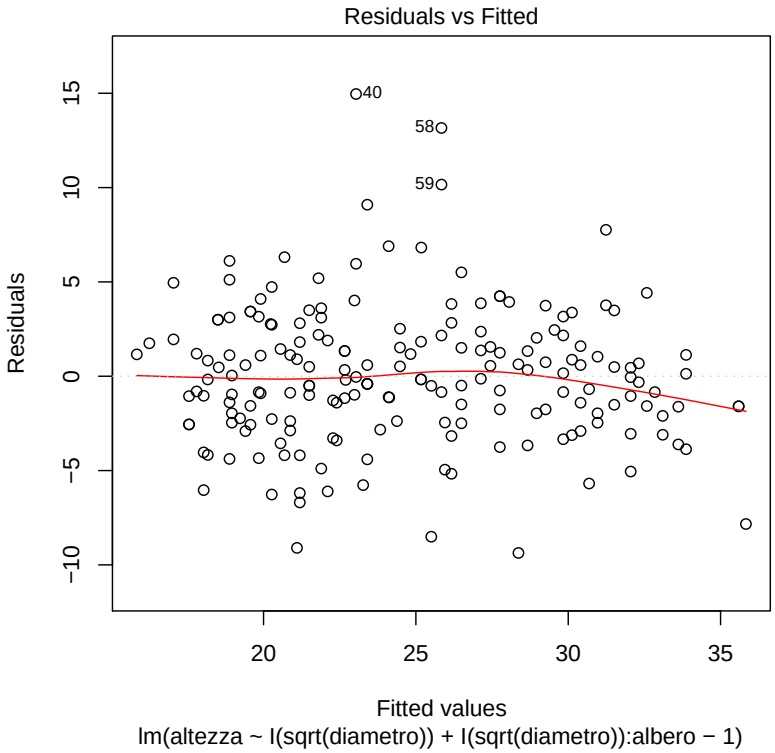
```
> summary(modello.d)

Call:
lm(formula = altezza ~ I(sqrt(diametro)) + I(sqrt(diametro)):albero -
    1)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3660 -2.2599 -0.1680  2.0163 14.9628

Coefficients:
                          Estimate Std. Error
I(sqrt(diametro))          4.13760    0.04824
I(sqrt(diametro)):albero -0.50353    0.09909
                          t value Pr(>|t|)
I(sqrt(diametro))          85.767   <2e-16 ***
I(sqrt(diametro)):albero   -5.082    9e-07 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.639 on 188 degrees of freedom
Multiple R-squared: 0.9798,        Adjusted R-squared: 0.9796
F-statistic:  4559 on 2 and 188 DF,  p-value: < 2.2e-16
```

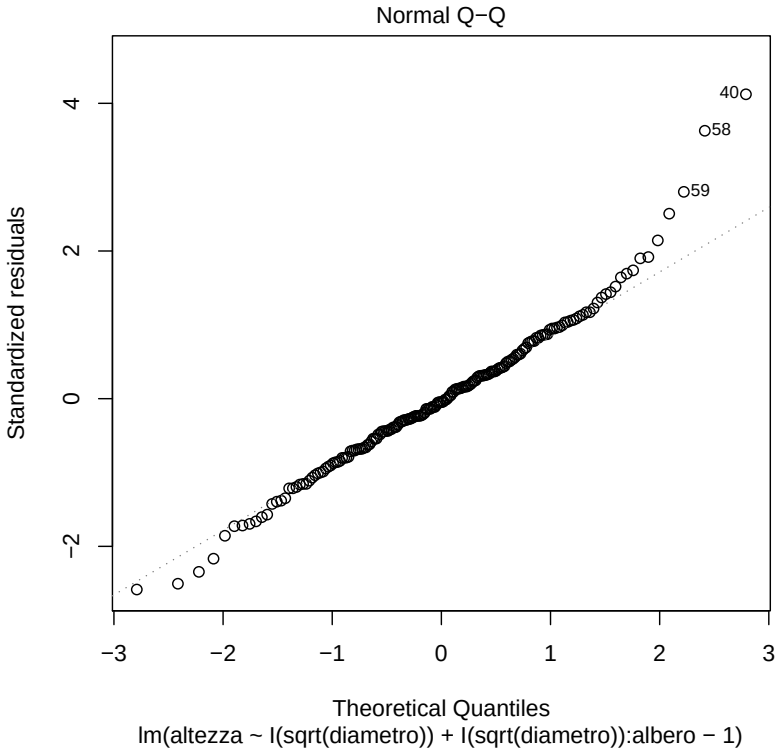
Default plot for the second model

```
> plot(modello.d, which = 1)

> plot(modello.d, which = 2)

> plot(modello.d, which = 3)

> plot(modello.d, which = 5)
```

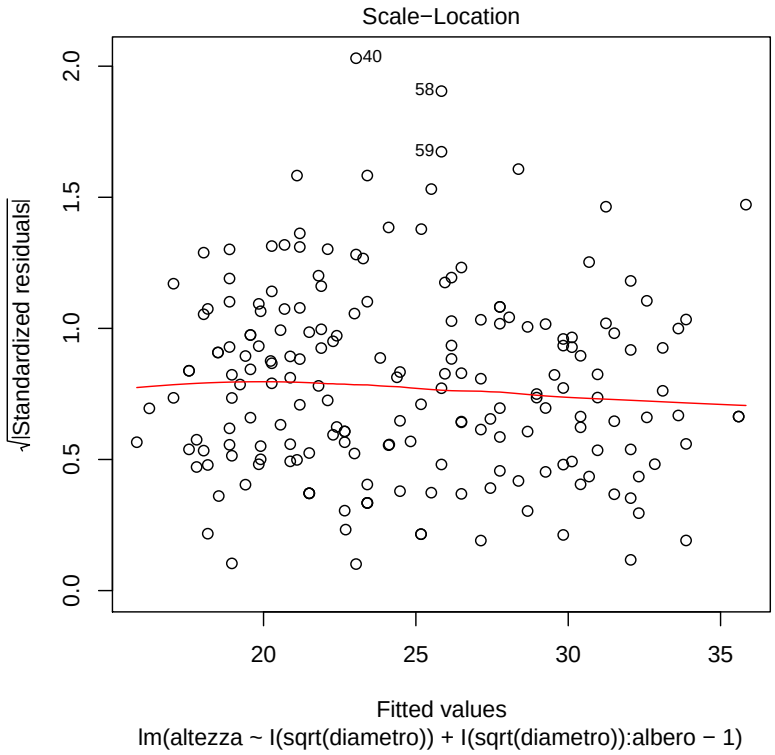
Residuals vs Fitted Values



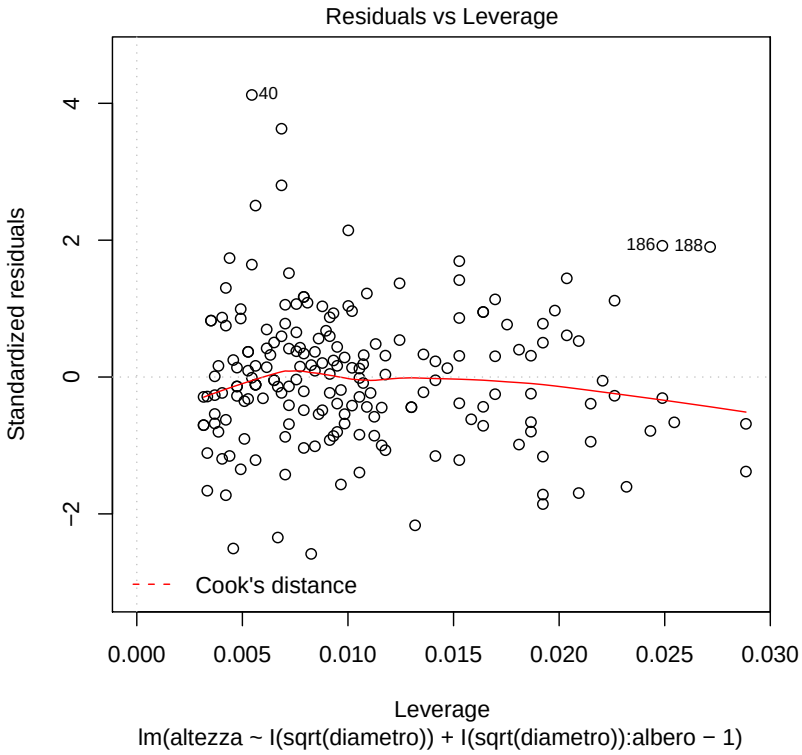
Standardized Residuals vs Theoretical Quantiles



Squared root of Standardized Residuals vs Fitted Values

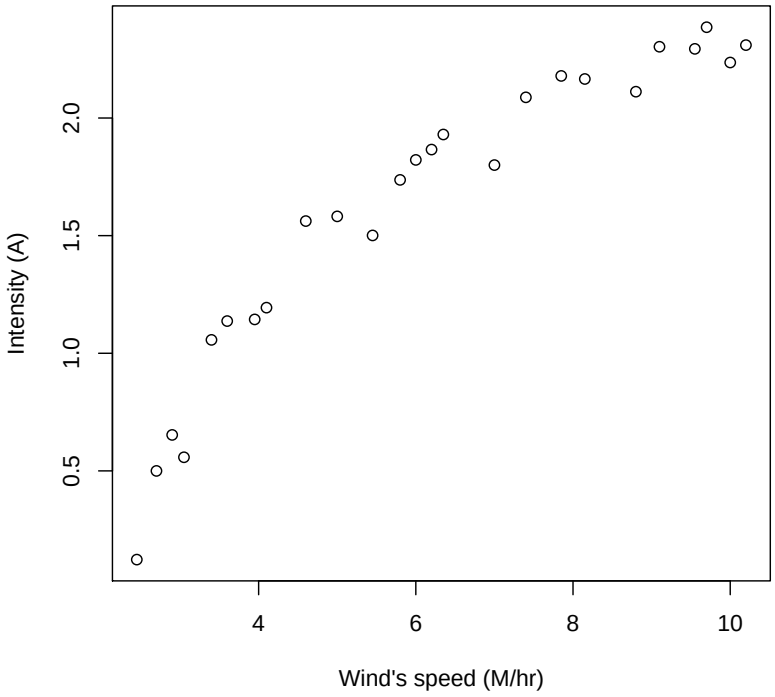


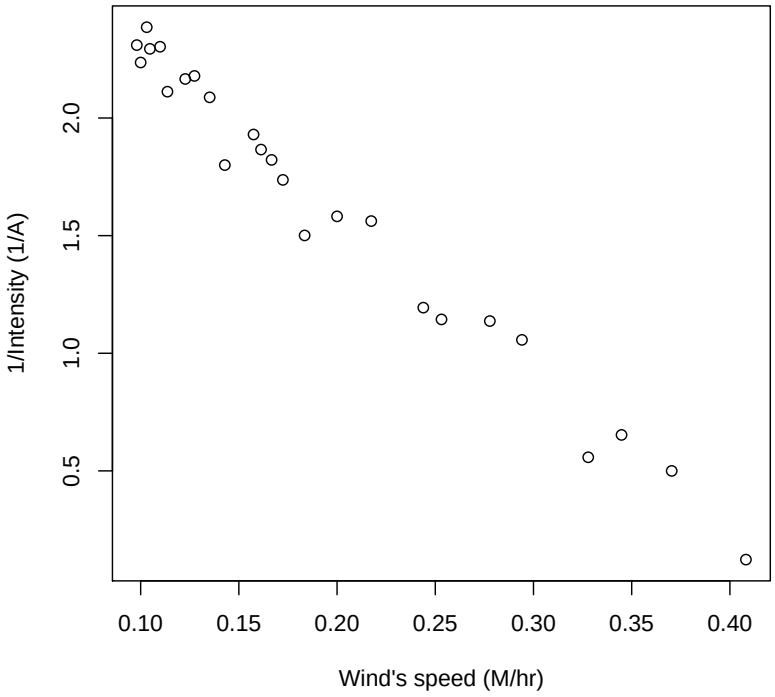
Standardized Residuals vs Leverage Points



Aeolian Energy

This dataset consists of 25 observations on two variables: Y, the Intensity (A) and X the speed of the wind (M/hr). We are looking for a good way of model the relationship between these two variables.


Intensity (A)
0.5
1.0
1.5
2.0
4
6
8
10
Wind's speed (M/hr)

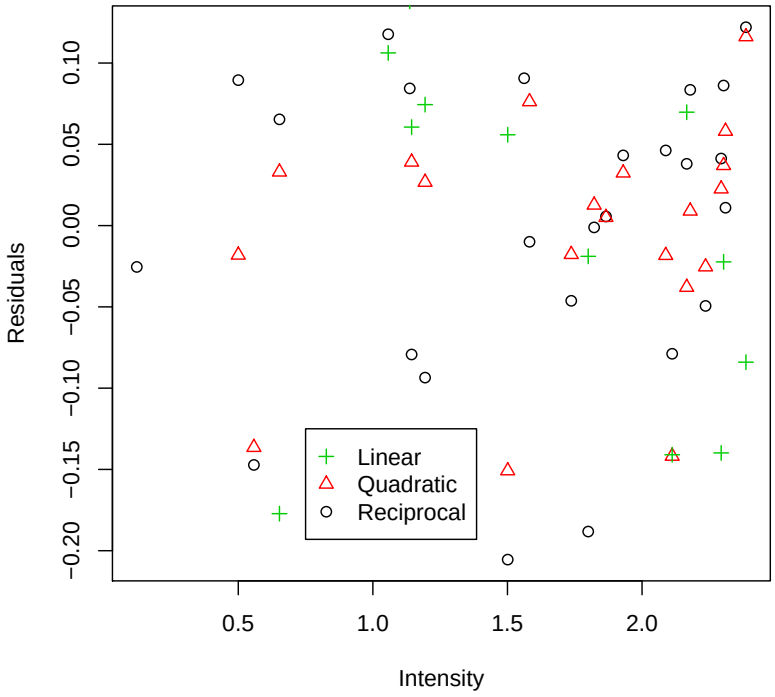


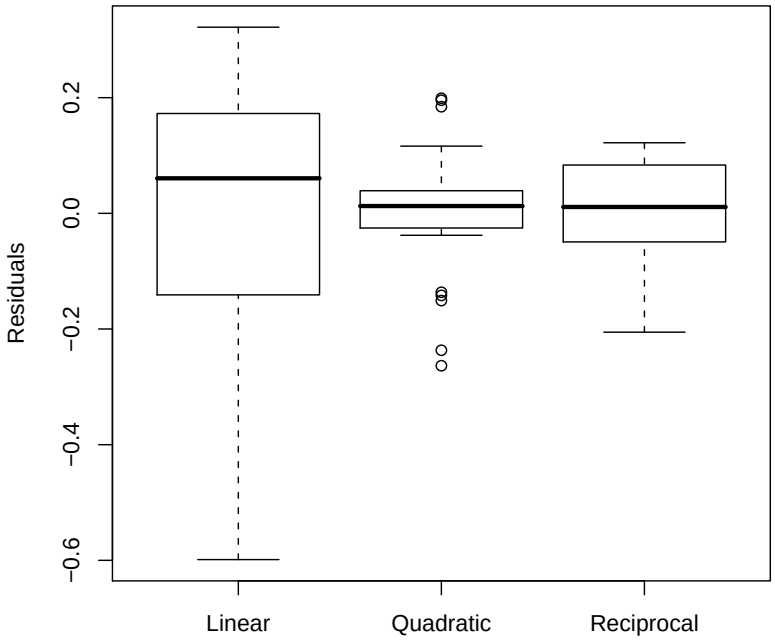
Three different models

Given X the wind's speed and Y the intensity, let us consider the following three models

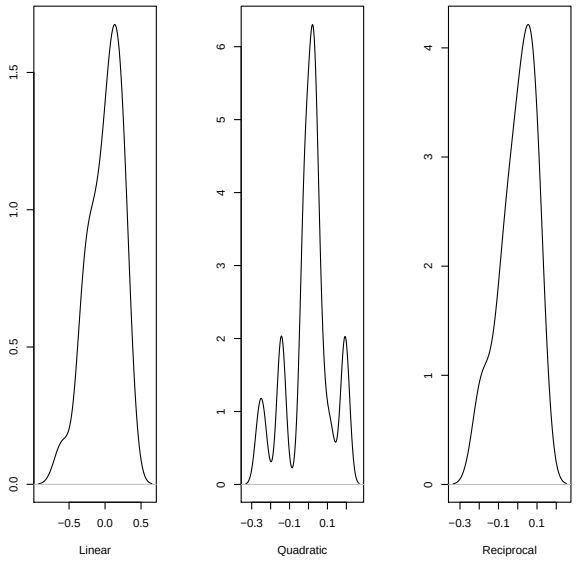
$$\begin{aligned} \text{Linear} Y &= \alpha + \beta X + \varepsilon \\ \text{Quadratic} Y &= \alpha + \beta X^2 + \varepsilon \\ \text{Reciprocal} Y &= \alpha + \beta \frac{1}{X} + \varepsilon \end{aligned}$$

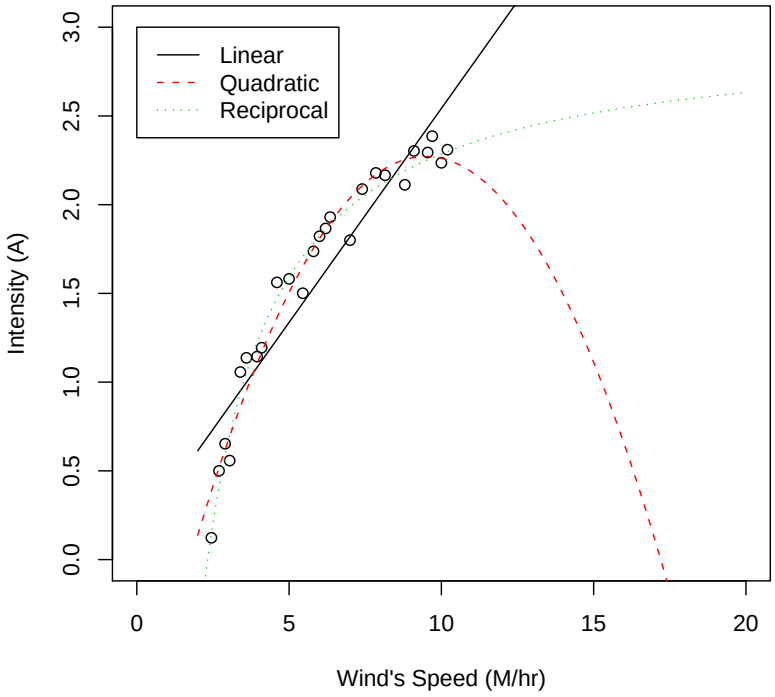
	R^2
Linear	0.874
Quadratic	0.968
Reciprocal	0.980


Residuals
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
0.5
1.0
1.5
2.0
Intensity
Linear
Quadratic
Reciprocal



Nonparametric Kernel Density Estimator of the Residuals





Hald Dataset

This (famous) dataset is an experiment on the composition of cement. There are four possible explanatory variables and one dependent variable:

X_1 *tricalcium aluminate*

X_2 *tricalcium silicate*

X_3 *tetracalcium alumino ferrite*

X_4 *dicalcium silicate*

Y *heat evolved* (Cal/gr).

The dataset is presented (among others) in Montgomery and Peck [1982].

Y	X_1	X_2	X_3	X_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

First Question

For each X_i let us consider the following model

$$Y = \alpha_0 + \alpha_1 \; X_i + \alpha_2 \; X_i^2 + \cdots + \alpha_{k-1} \; X_i^{k-1} + \varepsilon$$

where $k-1$ is the maximum degree of the polynomial such that the corresponding matrix has full rank.

1. For each couple (X_i, Y) indicate the value of k.
2. Without estimate $\hat{\alpha}_j$, $(j = 0, \cdots, k-1)$ indicate for wich X_i the residual deviance ${\delta_{k-1}}^2$ reach the minimum value.

To answer this first question let us count the number of all linear indipendent observations, which are equal in this case, to the number of distinct observations

	k
X_1	7
X_2	12
X_3	9
X_4	11

To answer to the second question we have to evaluate η^2. The mean of Y is $\bar{Y} = \mu_Y = 95.423$ and the variance is $\sigma_Y^2 = 208.905$. While the observations are [1mm]

i	$X_1 = x_{1i}$	Y				$\mu_{Y\mid\ X_1=x_{1i}}$	f_i
1	1	74.3	72.5	83.8		76.867	3
2	2	93.1				93.100	1
3	3	102.7				102.700	1
4	7	78.5	95.9			87.200	2
5	10	109.4				109.400	1
6	11	104.3	87.6	109.2	113.3	103.600	4
7	21	115.9				115.900	1

from which the variance of the conditional mean is

$$\sigma^2_{Y\mid\ X_1} = \frac{\sum_{i=1}^{k}(\mu_{Y\mid\ X_1=x_{1i}} - \mu_Y)^2\ f_i}{\sum_{i=1}^{k} f_i} = 162.209$$

For the second variable

i	$X_2 = x_{2i}$	Y		$\mu_{Y\mid\ X_2=x_{2i}}$	f_i
1	26	78.5		78.500	1
2	29	74.3		74.300	1
3	31	87.6	72.5	80.050	2
4	40	83.8		83.800	1
5	47	115.9		115.900	1
6	52	95.9		95.900	1
7	54	93.1		93.100	1
8	55	109.2		109.200	1
9	56	104.3		104.300	1
10	66	113.3		113.300	1
11	68	109.4		109.400	1
12	71	102.7		102.700	1

and the conditional variance is $\sigma^2_{Y\mid\ X_2} = 200.135$

For the third variable

i	$X_3 = x_{3i}$	Y			$\mu_{Y\mid\ X_3=x_{3i}}$	f_i
1	4	115.9			115.900	1
2	6	78.5	95.9		87.200	2
3	8	104.3	87.6	109.4	100.433	3
4	9	109.2	113.3		111.250	2
5	15	74.3			74.300	1
6	17	102.7			102.700	1
7	18	93.1			93.100	1
8	22	72.5			72.500	1
9	23	83.8			83.800	1

and the conditional variance is $\sigma^2_{Y|\ X_3} = 176.610$.

For the fourth variable

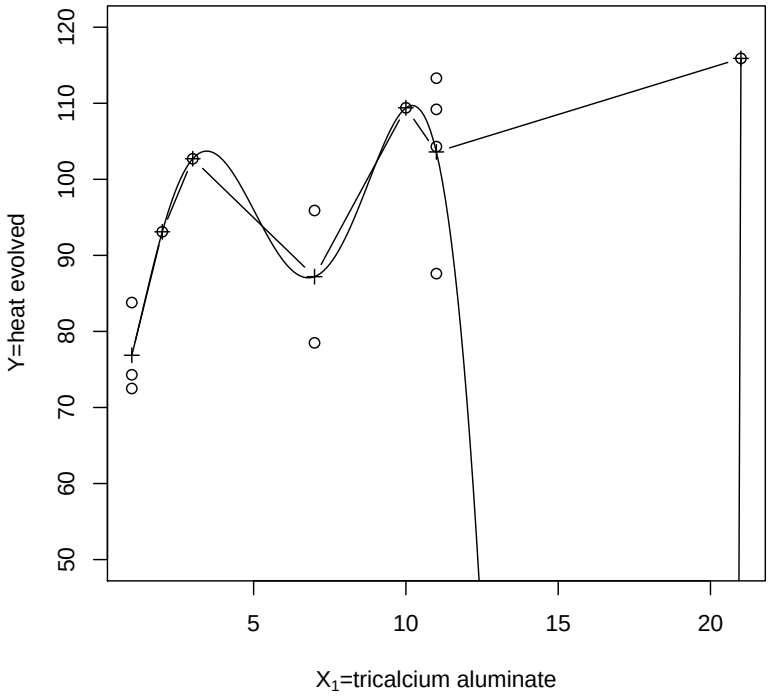
i	$X_4 = x_{4i}$	Y		$\mu_{Y|\ X_4=x_{4i}}$	f_i
1	6	102.7		102.700	1
2	12	113.3	109.4	111.350	2
3	20	104.3		104.300	1
4	22	109.2	93.1	101.150	2
5	26	115.9		115.900	1
6	33	95.9		95.900	1
7	34	83.8		83.800	1
8	44	72.5		72.500	1
9	47	87.6		87.600	1
10	52	74.3		74.300	1
11	60	78.5		78.500	1

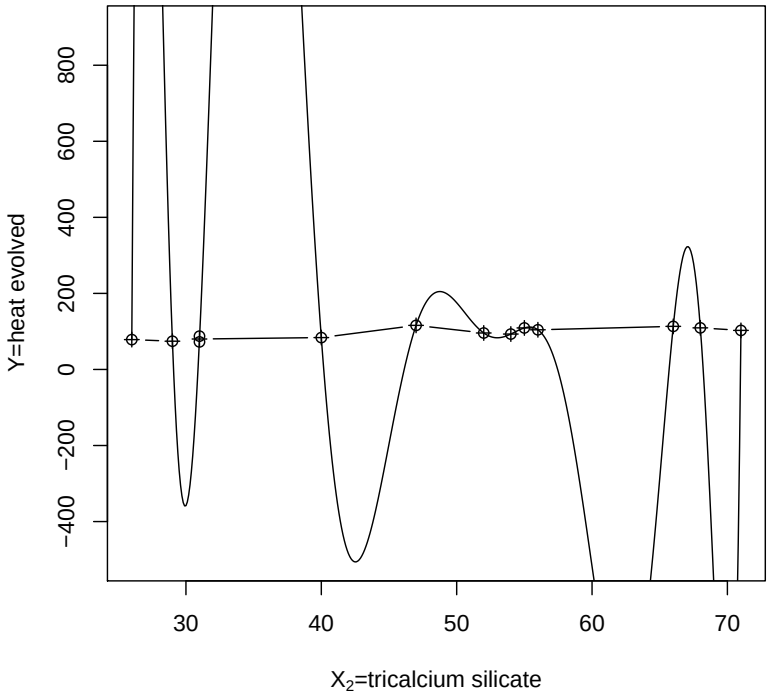
and the conditional variance is $\sigma^2_{Y|\ X_4} = 198.350$.

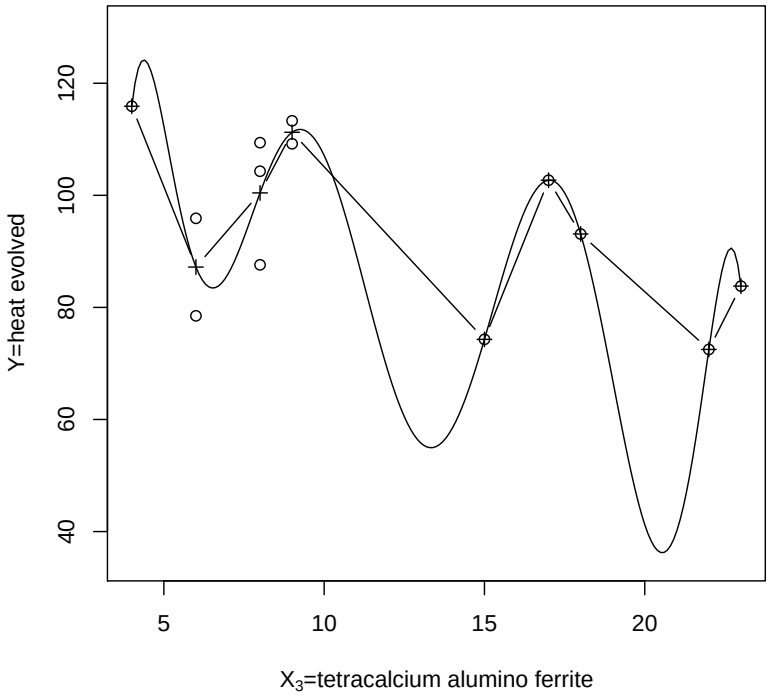
Hence $\eta^2_{YX_i} = \frac{\sigma^2_{Y|\ X_i}}{\sigma^2_Y}$ for the four different models are

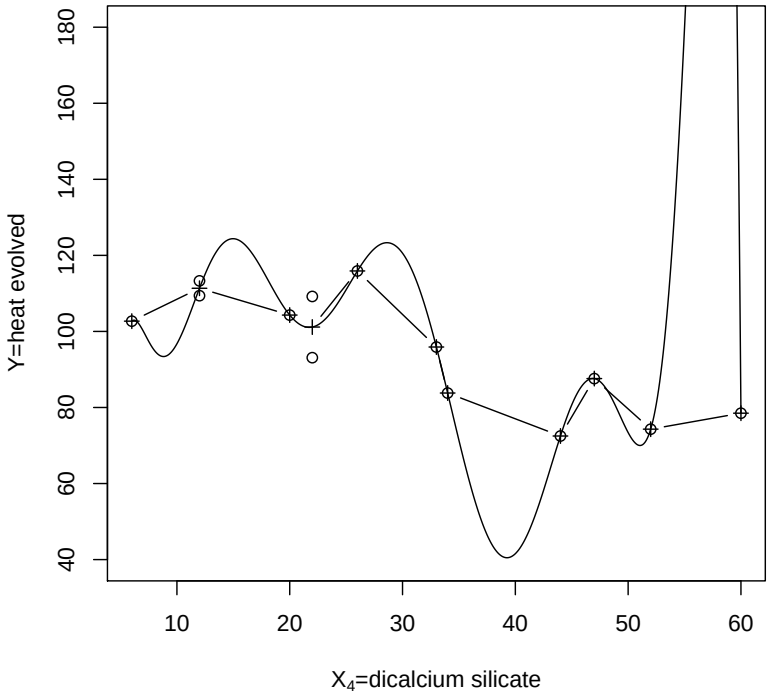
$$\begin{array}{lll} \eta^2_{YX_1} & 162.209/208.905 & 0.776 \\ \eta^2_{YX_2} & 200.135/208.905 & 0.958 \\ \eta^2_{YX_3} & 176.610/208.905 & 0.845 \\ \eta^2_{YX_4} & 198.350/208.905 & 0.949 \end{array}$$

then the best model (since η^2 criterion) is based on the X_2 polynomial.


Y=heat evolved
50
60
70
80
90
100
110
120
5
10
15
20
X1=tricalcium aluminate


Y=heat evolved
800
600
400
200
0
−200
−400
30
40
50
60
70
X2=tricalcium silicate


Y=heat evolved
40
60
80
100
120
5
10
15
20
X3=tetracalcium alumino ferrite



Second Question

Let us consider models on the form

$$Y = \alpha + \beta X_i + \varepsilon \qquad i = 1, \cdots, 4 \ .$$

Find out a criterion to select the explanatory variable X_i in order to explain the variability of Y without estimate α and β.

In this situation we may use the Pearson Correlation Coefficient between Y and X_i

$$\begin{aligned} \rho_{YX_1} &= \ \ 0.731 \\ \rho_{YX_2} &= \ \ 0.816 \\ \rho_{YX_3} &= -0.535 \\ \rho_{YX_4} &= -0.821 \end{aligned}$$

and hence the best model (since ρ) is based on X_4.

Third Question

Choose the best model based on the couple (X_1, X_j) $(j = 2, \cdots, 4)$ in order to explain the variability of Y. That is, let us consider models on the

form

$$Y = \alpha + \beta X_1 + \gamma X_j + \varepsilon \qquad i = 2, \cdots, 4 \ .$$

To answer this question we have to consider the Partial Correlation Coefficient which is the Pearson Correlation Coefficient between the regression residuals of two models with different response variables and the same explanatory form. In our case, let

$$r_Y = Y - \hat{\alpha}_Y - \hat{\beta}_Y X_1$$

the residual of the regression of Y with respect to (only) X_1 and

$$r_{X_j} = X_j - \hat{\alpha}_{X_j} - \hat{\beta}_{X_j} X_1 \qquad j = 2, \cdots, 4$$

the residual of the regression of X_j with respect to (only) X_i then

$$\rho_{Y,X_j|X_1} = \rho_{r_Y, r_{X_j}} \qquad j = 2, \cdots, 4 \ .$$

$$\begin{array}{rcr} \rho_{Y,X_2|X_1} & = & 0.977 \\ \rho_{Y,X_3|X_1} & = & 0.175 \\ \rho_{Y,X_4|X_1} & = & \text{-}0.970 \end{array}$$

so that, the best model is based on the couple (X_1, X_2) between the considered models.

Fourth Question

Why in the last question the included variable is X_2 instead of X_4?

The answer is linked to the correlation structure of the explanatory variables:

	X_1	X_2	X_3	X_4
X_1	1.000	0.229	-0.824	-0.245
X_2	0.229	1.000	-0.139	-0.973
X_3	-0.824	-0.139	1.000	0.029
X_4	-0.245	-0.973	0.029	1.000

1 (In)Dependence

1.1 Type of Indepencence

Three different way of measure independence between random variables

In statistics, related to the linear models, there are three different way of measuring the (in)dependence between random variables:

1. Stochastic independence;
2. Independence in mean;
3. Linear independence:

We will discuss these three type mainly in a 2 dimension setting, the extension to the general dimension is briefly discussed or left to the reader.

Context

Let X and Y two continuous random variables with density $f_X(x)$ and $f_Y(y)$ (respectively) with respect to some dominating measure ν and distribution functions $F_X(x)$ and $F_Y(y)$. Their join density is $f_{X,Y}(x,y)$ with distribution function $F_{X,Y}(x,y)$.

Definition 165 (Stochastic Independence) Two random variables X and Y are said **stochastic independent** if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \qquad \forall (x,y) \in \mathbb{R}^2 \ .$$

If the condition does not hold then they are **stochastic dependent**. △

This is a symmetric concept. That is, if X is independent from Y then Y is independent from X.

Homework 166 The definition is equivalent to the following statements:

1. $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all $(x,y) \in \mathbb{R}^2$;
2. $f_{X|Y(x|y)} = f_X(x)$ for all $(x,y) \in \mathbb{R}^2$;
3. $f_{Y|X(y|x)} = f_Y(y)$ for all $(x,y) \in \mathbb{R}^2$; ⊠

Example 167 Let $(X,Y) \sim f_{X,Y}(x,y)$ with $f_{X,Y}(x,y) = y\lambda \exp(-(\lambda+x)y)$ and $(x,y) \in \mathbb{R}^+ \times \mathbb{R}^+$. The marginal density of Y is

$$\begin{aligned} f_Y(y) &= \int_0^{+\infty} f_{X,Y}(x,y) d\nu(x) \\ &= \int_0^{+\infty} y\lambda \exp(-(\lambda+x)y)\, d\nu(x) \\ &= \lambda \exp(-\lambda y) \int_0^{+\infty} y \exp(-yx)\, d\nu(x) \\ &= \lambda \exp(-\lambda y) \end{aligned}$$

that is $Y \sim \text{Exp}(\lambda)$. Hence

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = y \exp(-yx)$$

that is, $X|Y = y \sim \text{Exp}(y)$. Since the conditional distribution is a function of y then X and Y are stochastic dependent. ⊠

Definition 168 (Conditional Mean) The **conditional mean** of $Y|X$ is

$$\text{E}(Y|X=x) = \int y f_{Y|X}(y|x) d\nu(y)$$

and it is a function of x over its domain. △

Theorem 169 Given a function $h(X,Y)$ (which is a r.v.) then

$$\text{E}(h(X,Y)) = \text{E}(\text{E}(h(X,Y)|X)) = \text{E}(\text{E}(h(X,Y)|Y)) \qquad \square$$

Definition 170 (Independence in Mean) The random variable Y is independent in mean from X if

$$\text{E}(Y|X=x) = \text{E}(Y) \qquad \forall x\ . \qquad \triangle$$

This is an asymmetric concept. That is, Y could be independent in mean from X but X could be depedent in mean from Y.

Example 171 (Cont.) The conditional mean of the random variable $X|Y = y$ is

$$\text{E}(X|Y=y) \int_0^{+\infty} y \exp(-yx)\, d\nu(x) = \frac{1}{y}$$

and since this is a function of y then X is dependent in mean from Y. ⊠

Homework 172 Is Y dependent in mean from X? What is $\mathrm{E}(Y|X=x)$?⊠

Homework 173 Let $\mathrm{var}(X|Y=y)$ be the conditional variance of X given $Y=y$. X is **independent in variance** from Y if

$$\mathrm{var}(X|Y=y) = \mathrm{var}(X) \qquad \forall y \ .$$

In the previous example, is X independent in variance from Y? What is $\mathrm{var}(X|Y=y)$? ⊠

Theorem 174 (Relation between Stochastic and In Mean Independence) If X and Y are stochastic independent then

1. X is independent in mean from Y;
2. Y is independent in mean from X; □

Example 175 The vice versa may not hold. Here a counterexample. Let X, Y be two discrete random variables with join probability distribution

		Y		
		-1	0	1
	-1	1/5	0	1/5
X	0	0	1/5	0
	1	1/5	0	1/5

The two random variables are independent in mean each others, however they are not stochastic independent as it is easy to verify. ⊠

Homework 176 Prove that if X and Y are stochastic independent then

1. X is independent in variance from Y;
2. Y is independent in variance from X; ⊠

Proof Since X and Y are stochastic independent then $f_{X|Y}(x|y) = f_X(x)$ for all $y \in \mathbb{R}$ then

$$\mathrm{E}(X|Y=y) = \int x f_{X|Y}(x|y) d\nu(x) = \int x f_X(x) d\nu(x) = \mathrm{E}(X) \quad y \in \mathbb{R}$$

that is, X is independent in mean from Y. In the same way we can show that Y is independent in mean from X. ■

Definition 177 (Bravais–Pearson Index of Correlation) The correlation between two random variables X and Y is

$$\operatorname{cor}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} \qquad \triangle$$

Homework 178 Using Cauchy–Schwarz Inequality prove that $-1 \leq \operatorname{cor}(X,Y) \leq 1$. ⊠

Definition 179 (Linear Independence) X and Y are **linear independent** if

$$\operatorname{cor}(X,Y) = 0 \qquad \triangle$$

Which is equivalent to $\operatorname{cov}(X,Y) = 0$. This is a symmetric concept.

Example 180 Let $X \sim \text{Unif}[-1,1]$ and $Y = X^2$. Hence, $f_X(x) = \frac{1}{2}\,\mathrm{I}_{[-1,1]}(x)$ and $f_{Y|X=x}(y|x) = \mathrm{I}_{[x^2]}(y)$. The join density is

$$f_{X,Y}(x,y) = \frac{1}{2}\,\mathrm{I}_{[-1,1]}(x)\,\mathrm{I}_{[x^2]}(y) \qquad x \in [-1,1] \quad y \in [0,1]\ .$$

Since $\mathrm{E}(X) = 0$ then $\operatorname{cov}(X,Y) = \mathrm{E}(XY)$ and

$$\begin{aligned}
\mathrm{E}(XY) &= \int xy f_{X,Y}(x,y) d\nu(x) d\nu(y) \\
&= \frac{1}{2}\int_{-1}^{1} x\,\mathrm{I}_{[-1,1]}(x) \int_0^1 y\,\mathrm{I}_{[x^2]}(y) d\nu(y) d\nu(x) \\
&= \frac{1}{2}\int_{-1}^{1} x\ x^2 d\nu(x) \\
&= \frac{1}{2}\int_{-1}^{1} x^3 d\nu(x) \\
&= \frac{1}{8} x^4|_{-1}^{1} d\nu(x) \\
&= 0
\end{aligned}$$

Hence X and Y are uncorrelated. ⊠

Homework 181 Are X and Y stochastic independent? Are X and Y independent in mean? ⊠

Theorem 182 (Relation between Linear and In Mean Independence) If X is independent in mean from Y then X and Y are linear independent.□

Proof Since X is independent in mean from Y then $\mathrm{E}(X|Y=y) = \mathrm{E}(X)$ for all $y \in \mathbb{R}$. Then

$$\begin{aligned}
\mathrm{E}(XY) &= \int\int xyf_{X|Y}(x|y)f_Y(y)d\nu(x)d\nu(y) \\
&= \int yf_Y(y)\int xf_{X|Y}(x|y)d\nu(x)d\nu(y) \\
&= \int yf_Y(y)\,\mathrm{E}(X|Y=y)d\nu(y) \\
&= \int yf_Y(y)\,\mathrm{E}(X)d\nu(y) \qquad \text{Independence in Mean} \\
&= \mathrm{E}(X)\,\mathrm{E}(Y)
\end{aligned}$$

and hence

$$\begin{aligned}
\mathrm{cov}(X,Y) &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) \\
&= \mathrm{E}(X)\,\mathrm{E}(Y) - \mathrm{E}(X)\,\mathrm{E}(Y) = 0
\end{aligned}$$ ■

Theorem 183 For the random vector (X,Y) we have

$$\mathrm{cov}(X,Y) = \mathrm{cov}(X,\mathrm{E}(Y|X)) = \mathrm{cov}(\mathrm{E}(X|Y),Y)$$ □

Proof

$$\begin{aligned}
\mathrm{cov}(X,Y) &= \mathrm{E}\left((X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right) \\
&= \mathrm{E}\left[\mathrm{E}\left((X-\mathrm{E}(X))(Y-\mathrm{E}(Y))|X\right)\right] \\
&= \mathrm{E}\left[(X-\mathrm{E}(X))\,\mathrm{E}\left((Y-\mathrm{E}(Y))|X\right)\right] \\
&= \mathrm{E}\left[(X-\mathrm{E}(X))(\mathrm{E}(Y|X)-\mathrm{E}(Y))\right] \\
&= \mathrm{cov}(X,\mathrm{E}(Y|X))
\end{aligned}$$ ■

Homework 184 Prove that $\mathrm{E}_X(E(Y|X)) = E(Y)$. ⊠

1.2 Decomposition of the Variance

Definition 185 (Regression function) Given a set of points $\boldsymbol{x} = (x_1, \cdots, x_n)$ any function $g(x)$ such that

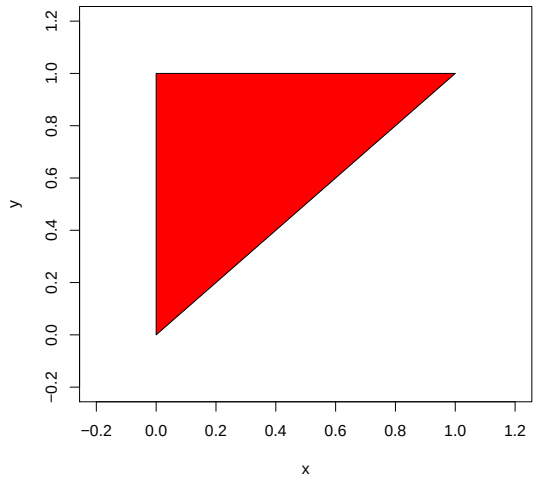
$$g(x) = \mathrm{E}(Y|X=x) \qquad \forall x \in \boldsymbol{x}$$

is called a **regression function** of $Y|X$ for the set $\boldsymbol{x}$. △

When instead a finite number of points $\boldsymbol{x}$ we consider the set $\mathcal{X} = \mathbb{R}$ then we must have $g(X) = \mathrm{E}(Y|X)$ with probability 1.

Example 186 Given the random vector (X, Y) with join density

$$f_{X,Y}(x,y) = \begin{cases} 2 & 0 \leq x \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$



Support of the join density $f_{X,Y}(x,y)$.

The marginal densities are

$$f_X(x) = \begin{cases} \int_x^1 2 d\nu(y) = 2(1-x) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Y(y) = \begin{cases} \int_0^y 2 d\nu(x) = 2y & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

While the conditional densities are

$$f_{X|Y}(x|y) = \begin{cases} \frac{2}{2y} = \frac{1}{y} & 0 \leq x \leq y \ , \ 0 < y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

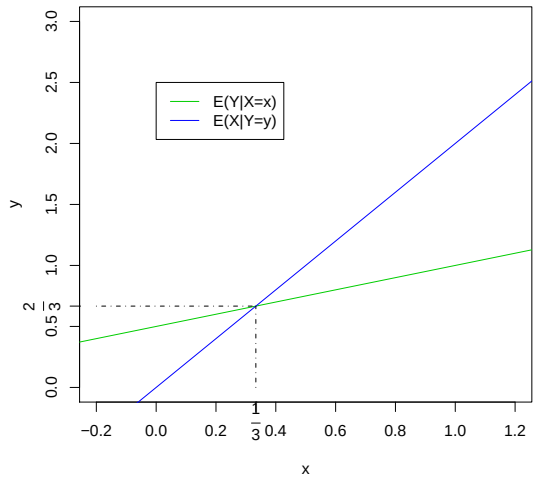
$$f_{Y|X}(y|x) = \begin{cases} \frac{2}{2(1-x)} = \frac{1}{1-x} & 0 \leq x < 1 \ , \ x \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Hence the conditional mean functions (and the regression functions) are

$$\begin{aligned}\mathrm{E}(Y|X=x) &= \int_x^1 \frac{1}{1-x} y d\nu(y) \\ &= \frac{1}{1-x}\frac{1-x^2}{2} \\ &= \frac{1}{2}+\frac{1}{2}x \qquad 0 \le x < 1\end{aligned}$$

and

$$\mathrm{E}(X|Y=y) = \int_0^y x\frac{1}{y} d\nu(x) = \frac{1}{2}y \qquad 0 < y \le 1$$



Regression function of Y given X (green) and X given Y (blue). ⊠

Theorem 187 Given a random vector (X, Y) such that $\mathrm{var}(Y)$ exists and a function $h(x)$ such that $h(X)$ is a random variable wih finite variance. Then $\mathrm{E}(Y - h(X))^2$ has a minimum when $h(x) = \mathrm{E}(Y|X = x)$. □

Proof

$$\mathrm{E}(Y - h(X))^2 = \mathrm{E}\left(\mathrm{E}\left((Y - h(X))^2|X\right)\right)$$

but for a fixed value of X the conditional expectation is

$$\begin{aligned}\mathrm{E}\left((Y - h(X))^2|X = x\right) &= \int (y - h(x))^2 f_{Y|X}(y|x) d\nu(y) \\ &\ge \int (y - \mathrm{E}(Y|X=x))^2 f_{Y|X}(y|x) d\nu(y) \qquad (1)\end{aligned}$$

for the property of the variance

and hence

$$\begin{aligned}\mathrm{E}(Y-h(X))^2 &\geq \int f_X(x)\int (y-\mathrm{E}(Y|X=x))^2 f_{Y|X}(y|x)d\nu(y)d\nu(x)\\ &\geq \int\int (y-\mathrm{E}(Y|X=x))^2 f_{X,Y}(x,y)d\nu(y)d\nu(x)\end{aligned}$$

so that

$$\mathrm{E}(Y-h(X))^2 \geq \mathrm{E}(Y-\mathrm{E}(Y|X))^2 \ . \tag{2}$$

The equality in the last formula holds only if the equality holds in formula (1). But we have

$$\begin{aligned}\int (y-h(x))^2 f_{Y|X}(y|x)d\nu(y) = &\int (y-\mathrm{E}(Y|X=x))^2 f_{Y|X}(y|x)d\nu(y)\\ &+\mathrm{E}\left((h(x)-\mathrm{E}(Y|X))^2\right)\end{aligned}$$

and so equality holds in formula (2) only if $h(X)=\mathrm{E}(Y|X)$ with probability one, that is $h(x)$ is a regression function. ■

Theorem 188 (Decomposition of the variance) Given two random variables X and Y then

$$\mathrm{var}(Y) = \mathrm{var}_X\left(\mathrm{E}(Y|X=x)\right) + \mathrm{E}_X(\mathrm{var}(Y|X=x))$$

where $\mathrm{var}(Y|X=x)$ is the conditional variance (function) or simply the variance of the random variable $Y|X=x$ that is

$$\begin{aligned}\mathrm{var}(Y|X=x) &= \int \left(y-\mathrm{E}(Y|X=x)\right)^2 f_{Y|X}(y|x)d\nu(y)\\ &= \mathrm{E}\left[(Y|X=x)^2\right] - \left[\mathrm{E}(Y|X=x)\right]^2 \ .\end{aligned}$$ □

- $\mathrm{var}_X\left(\mathrm{E}(Y|X=x)\right)$ is called the **between** (or explained) variance;
- $\mathrm{E}_X(\mathrm{var}(Y|X=x))$ is called the **within** (or residual) variance.

Example 189

Proof

$$\begin{aligned}
\mathrm{var}(Y) &= \int (y - \mathrm{E}(Y))^2 f_Y(y) d\nu(y) \\
&= \int (y - \mathrm{E}(Y))^2 \int f_{X,Y}(x,y) d\nu(x) d\nu(y) \\
&= \int \int (y - \mathrm{E}(Y))^2 f_{X,Y}(x,y) d\nu(x) d\nu(y) \\
&= \int \int (y - \mathrm{E}(Y|X=x))^2 f_{X,Y}(x,y) d\nu(x) d\nu(y) \\
&+ \int \int (\mathrm{E}(Y|X=x) - \mathrm{E}(Y))^2 f_{X,Y}(x,y) d\nu(x) d\nu(y) \\
&+ 2 \int \int (y - \mathrm{E}(Y|X=x)) (\mathrm{E}(Y|X=x) - \mathrm{E}(Y)) f_{X,Y}(x,y) d\nu(x) d\nu(y)
\end{aligned}$$

but the last term is zero since

$$\begin{aligned}
&= \int \int (y - \mathrm{E}(Y|X=x)) (\mathrm{E}(Y|X=x) - \mathrm{E}(Y)) f_{X,Y}(x,y) d\nu(x) d\nu(y) \\
&= \int (\mathrm{E}(Y|X=x) - \mathrm{E}(Y)) \int (y - \mathrm{E}(Y|X=x)) f_{Y|X}(y|x) d\nu(y) f_X(x) d\nu(x)
\end{aligned}$$

and

$$\begin{aligned}
&= \int (y - \mathrm{E}(Y|X=x)) f_{Y|X}(y|x) d\nu(y) \\
&= \int y f_{Y|X}(y|x) d\nu(y) - \mathrm{E}(Y|X=x) \\
&= 0 \ .
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{var}(Y) &= \int \int (y - \mathrm{E}(Y|X=x))^2 f_{Y|X}(y|x) d\nu(y) f_X(x) d\nu(y) \\
&+ \int \int (\mathrm{E}(Y|X=x) - \mathrm{E}(Y))^2 f_{Y|X}(y|x) d\nu(y) f_X(x) d\nu(x) \\
&= \int \mathrm{var}(Y|X=x) f_X(x) d\nu(x) \\
&+ \int (\mathrm{E}(Y|X=x) - \mathrm{E}(Y))^2 f_X(x) d\nu(x) \int f_{Y|X}(y|x) d\nu(y) \\
&= \mathrm{E}_X (\mathrm{var}(Y|X)) \\
&+ \mathrm{var}_X (\mathrm{E}(Y|X))
\end{aligned}$$

■

Definition 190 (Index of Dependence in Mean) The **index of dependence in mean** $\eta^2_{Y|X}$ is

$$\eta^2_{Y|X} = \frac{\text{var}_X(\text{E}(Y|X=x))}{\text{var}(Y)} \qquad \text{for } \text{var}(Y) > 0 \ .$$

The index is between 0 and 1. △

1. $\eta^2_{Y|X} = 0$ if and only if Y is independent in mean from X;
2. $\eta^2_{Y|X} = 1$ if and only if there exists a (deterministic) function $g(x)$ such that $\Pr(Y = g(X)) = 1$.

Theorem 191

- $\eta^2_{Y|X} = \text{cor}^2(Y, \text{E}(Y|X))$;
- $\eta^2_{X|Y} = \text{cor}^2(X, \text{E}(X|Y))$. □

Proof

$$\begin{aligned}
\text{cor}(Y, \text{E}(Y|X)) &= \frac{\text{cov}(Y, \text{E}(Y|X))}{\sqrt{\text{var}(Y)}\sqrt{\text{var}_X \text{E}(Y|X)}} \\
&= \text{E}\left(\frac{Y - \text{E}(Y)}{\sqrt{\text{var}(Y)}} \frac{\text{E}(Y|X) - \text{E}(Y)}{\sqrt{\text{var}_X(\text{E}(Y|X))}}\right) \\
&= \text{E}_X\left(\text{E}\left(\frac{Y - \text{E}(Y)}{\sqrt{\text{var}(Y)}} \frac{\text{E}(Y|X) - \text{E}(Y)}{\sqrt{\text{var}_X(\text{E}(Y|X))}}|X\right)\right) \\
&= \text{E}_X\left(\frac{\text{E}(Y|X) - \text{E}(Y)}{\sqrt{\text{var}_X(\text{E}(Y|X))}} \text{E}\left(\frac{Y - \text{E}(Y)}{\sqrt{\text{var}(Y)}}|X\right)\right) \\
&= \text{E}_X\left(\frac{(\text{E}(Y|X) - \text{E}(Y))^2}{\sqrt{\text{var}_X(\text{E}(Y|X))}\sqrt{\text{var}(Y)}}\right) \\
&= \frac{\text{var}_X(\text{E}(Y|X))}{\sqrt{\text{var}_X(\text{E}(Y|X))}\sqrt{\text{var}(Y)}} \\
&= \frac{\sqrt{\text{var}_X(\text{E}(Y|X))}}{\sqrt{\text{var}(Y)}} \geq 0
\end{aligned}$$

and hence

$$\text{cor}^2(Y, \text{E}(Y|X)) = \frac{\text{var}_X(\text{E}(Y|X))}{\text{var}(Y)} = \eta^2_{Y|X}$$ ■

Compare these results with those from Theorem 183.

Theorem 192 Given a random vector (X, Y) and a function $h(x)$ (such that $h(X)$ is a r.v.) and if $\mathrm{var}(Y)$ and $\mathrm{var}(h(X))$ exist positive then

$$0 \leq \mathrm{cor}^2(X, Y) \leq \max_h \mathrm{cor}^2(h(X), Y) = \mathrm{cor}^2(\mathrm{E}(Y|X), Y) = \eta^2_{Y|X} \leq 1 \quad \square$$

Proof From the very definition of the correlation coefficient

$$\begin{aligned}\mathrm{cor}^2(h(X), Y) &= \left(\mathrm{E}\left(\frac{h(X) - \mathrm{E}(h(X))}{\sqrt{\mathrm{var}(h(X))}} \frac{Y - \mathrm{E}(Y)}{\sqrt{\mathrm{var}(Y)}}\right)\right)^2 \\ &= \left(\mathrm{E}_X\left(\frac{h(X) - \mathrm{E}(h(X))}{\sqrt{\mathrm{var}(h(X))}} \mathrm{E}\left(\frac{Y - \mathrm{E}(Y)}{\sqrt{\mathrm{var}(Y)}}|X\right)\right)\right)^2\end{aligned}$$

but

$$\mathrm{E}\left(\frac{Y - \mathrm{E}(Y)}{\sqrt{\mathrm{var}(Y)}}|X\right) = \frac{\mathrm{E}(Y|X) - \mathrm{E}(Y)}{\sqrt{\mathrm{var}(Y)}}$$

and hence by Cauchy–Schwarz Inequality

$$\begin{aligned}\left(\mathrm{E}_X\left(\frac{h(X) - \mathrm{E}(h(X))}{\sqrt{\mathrm{var}(h(X))}} \frac{\mathrm{E}(Y|X) - \mathrm{E}(Y)}{\sqrt{\mathrm{var}(Y)}}\right)\right)^2 &\leq \mathrm{var}_X\left(\frac{h(X) - \mathrm{E}(h(X))}{\sqrt{\mathrm{var}(h(X))}}\right) \\ &\quad \times \mathrm{var}_X\left(\frac{\mathrm{E}(Y|X) - \mathrm{E}(Y)}{\sqrt{\mathrm{var}(Y)}}\right)\end{aligned}$$

that is

$$\mathrm{cor}^2(h(X), Y) \leq 1 \ \ \mathrm{var}_X\left(\frac{\mathrm{E}(Y|X) - \mathrm{E}(Y)}{\sqrt{\mathrm{var}(Y)}}\right) = \frac{\mathrm{var}_X(\mathrm{E}(Y|X))}{\mathrm{var}(Y)} = \eta^2_{Y|X} \quad \blacksquare$$

Theorem 193 Given a random vector (X, Y) then $\mathrm{cor}^2(X, Y) = \eta^2_{Y|X}$ if and only if the regression function $\mathrm{E}(Y|X = x)$ is linear. $\square$

Proof Let us assume that $\mathrm{E}(Y|X = x) = a + bx$ $(b \neq 0)$ is linear. Then

$$\eta^2_{Y|X} = \mathrm{cor}^2(\mathrm{E}(Y|X), Y) = \mathrm{cor}^2(a + bX, Y) = \mathrm{cor}^2(X, Y)$$

Let us assume that $\eta^2_{Y|X} = \mathrm{cor}^2(X, Y)$ since

$$\mathrm{cov}(X, Y) = \mathrm{cov}(X, \mathrm{E}(Y|X))$$

then

$$\begin{aligned}\mathrm{cor}^2(X,Y) &= \frac{\mathrm{cov}^2(X,\mathrm{E}(Y|X))}{\mathrm{var}(X)\,\mathrm{var}(Y)}\\ &= \frac{\mathrm{cov}^2(X,\mathrm{E}(Y|X))}{\mathrm{var}(X)\,\mathrm{var}_X(\mathrm{E}(Y|X))}\frac{\mathrm{var}_X(\mathrm{E}(Y|X))}{\mathrm{var}(Y)}\\ &\mathrm{cor}^2(X,\mathrm{E}(Y|X))\eta^2_{Y|X} \leq \eta^2_{Y|X}\end{aligned}$$

but $\mathrm{cor}^2(X,Y) = \eta^2_{Y|X}$ and hence $|\mathrm{cor}(X,\mathrm{E}(Y|X))| = 1$ that is, $\mathrm{E}(Y|X=x)$ must be a linear function. ■

2 Linear Models

2.1 Introduction

Definition 194 (Linear Model) Let $\boldsymbol{Z} = (Y, \boldsymbol{X})$ be a random vector so that Y be a random variable and $\boldsymbol{X}$ a random vector of length k. A **linear model** is such that

1. $Y|\boldsymbol{X} \sim N(\mathrm{E}(Y|\boldsymbol{X}), \sigma^2)$ (**Casual component**);
2. $\mathrm{E}(Y|\boldsymbol{X}=\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\beta}\rangle = \boldsymbol{x}^t\boldsymbol{\beta}$ (**Systematic component**).

where $\boldsymbol{\beta}$ is a vector of length k. △

- The first assumption states that:
  - The conditional distribution of $Y|\boldsymbol{X}$ is Normal;
  - The conditional variance of $Y|\boldsymbol{X}$ does not depend on the value of $\boldsymbol{X}$ (Assumption of **homoschedasticity**);
- The second assumption stated that the conditional mean of $Y|\boldsymbol{X}$ is a linear combination of the random variables X_i through the vector $\boldsymbol{\beta}$;
- No assumption is made on the joint distribution of $\boldsymbol{X}$.

Another formulation of the problem

Under a linear model we have

$$\varepsilon = Y - \mathrm{E}(Y|\boldsymbol{X}) \sim N(0, \sigma^2)$$

which is called the **error (stochastic) component**. Hence

$$\begin{aligned} Y &= \mathrm{E}(Y|\boldsymbol{X}) + \varepsilon \\ &= \boldsymbol{x}^t\boldsymbol{\beta} + \varepsilon \end{aligned}$$

which is the classical form for presenting linear models. From this formulation we can see that Y could be decomposed in a signal+noise form.

Linear model and a data sample

Let $Z = (\boldsymbol{y}, X)$ be an $n \times (k+1)$ matrix such that each row is a determination from the random vector $\boldsymbol{Z} = (Y, \boldsymbol{X})$ draw independently from the remains. That is

$$Z = (\boldsymbol{y}, X) = \begin{bmatrix} y_1 & x_{11} & x_{12} & \cdots & x_{1k} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} = \begin{bmatrix} y_1 & \boldsymbol{x}_1{}^t \\ y_2 & \boldsymbol{x}_2{}^t \\ \vdots & \vdots \\ y_n & \boldsymbol{x}_n{}^t \end{bmatrix}$$

Linear model and a data sample

This leads to n observational equations

$$y_i = \boldsymbol{x}_i{}^t\boldsymbol{\beta} + e_i \qquad 1 \leq i \leq n$$

that can be written in a matrix form as

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

where $\boldsymbol{e}$ is an n-vector containing the n unobserved values from the random variable ε. The elements of $\boldsymbol{e}$ are called **errors**. Since the sample scheme, the last form has a stochastic counterpart (keeping the values of X fixed)

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

where $\underset{n\times 1}{\boldsymbol{Y}}|X \sim N(\mathrm{E}(\boldsymbol{Y}|X), \sigma^2 I)$ and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I)$.

Summarizing

- $\boldsymbol{y}$ is a $n \times 1$ vector of observations from Y (or, one observation from the random vector $\boldsymbol{Y}$);

- X is an $n \times k$ matrix;
- $\boldsymbol{\beta}$ is an $p \times 1$ vector of the parameters;
- $\boldsymbol{e}$ is an $n \times 1$ vector of errors (or, one observation from the random vector $\boldsymbol{\varepsilon}$).

- Y is called the **dependent** variable (**endogenous** variable in the econometric literature);
- In linear model Y is always a quantitative, continuous r.v. (normally distributed);
- $\boldsymbol{X}$ is called the set of **independent** variables (**exogenous** variables in the econometric literature);
- Since the joint distribution of $\boldsymbol{X}$ is not relevant in many aspects of a linear model often X is considered as a deterministic matrix, in this case X is a **fixed carriers** otherwise is a **random carriers**;
- $\boldsymbol{X}$ may be a quantitative or a factor random vector.

Example 195 Given the three variables regarding the price of an used car:

- P: Price of the car;
- A: Age of the car (in month since the first sale);
- T: Type of the car (we restrict our attention to four different type of car, namely 0, 1, 2, 3).

Then using P as the dependent variable we have the following linear model

$$p_i = \beta_1 a_i + \beta_2 \, \mathrm{I}_{t_i=0} + \beta_3 \, \mathrm{I}_{t_i=1} + \beta_4 \, \mathrm{I}_{t_i=2} + \beta_5 \, \mathrm{I}_{t_i=3} + e_i$$

where $\mathrm{I}_{t_i=0}$ is the indicator variable with value 1 if $t_i = 0$ and 0 otherwise. In matrix form

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} a_1 & 1 & 0 & 0 & 0 \\ a_2 & 1 & 0 & 0 & 0 \\ a_3 & 0 & 1 & 0 & 0 \\ \cdots & & \cdots & & \\ a_i & 0 & 0 & 1 & 0 \\ \cdots & & \cdots & & \\ a_n & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_5 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

A linear model may have different parameterization, for instance the previous model could be rewritten as

$$p_i = \alpha_0 + \alpha_1 a_i + \alpha_2 \, \mathrm{I}_{t_i=1} + \alpha_3 \, \mathrm{I}_{t_i=2} + \alpha_4 \, \mathrm{I}_{t_i=3} + e_i$$

This parameterization is equivalent to the previous one, but the interpretation of the parameters $\boldsymbol{\alpha}$ is different. The matrix form is

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} a_1 & 1 & 0 & 0 & \\ a_2 & 1 & 0 & 0 & \\ a_3 & 1 & 1 & 0 & \\ \cdots & \cdots & & & \\ a_i & 1 & 0 & 1 & 0 \\ \cdots & \cdots & & & \\ a_n & 1 & 0 & 0 & \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_4 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Under the linear model we have specified, so far, the slope of the relation between the price and the age inside each type is identical. We may allow for different slope, for all or some types of cars, for instance the model

$$p_i = \gamma_0 + \gamma_1 a_i + \gamma_2 \, \mathrm{I}_{t_i=1} + \gamma_3 \, \mathrm{I}_{t_i=2} + \gamma_4 \, \mathrm{I}_{t_i=3} + \gamma_5 a_i \, \mathrm{I}_{t_i=3} + e_i$$

allows the type 3 to have a different slope with respect to the remains types. The term $a_i \, \mathrm{I}_{t_i=3}$ is called **interaction**. ⊠

2.2 Estimation using Least Squares and Likelihood

2.2.1 Introduction

A general procedure for the estimation of the parameters $\boldsymbol{\beta}$ is to minimize a suitable function of the errors:

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} M(e_i) = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} M(y_i - \boldsymbol{x}_i{}^t \boldsymbol{\beta}) \ .$$

Two examples are

1. $M(x) = |x|$;
2. $M(x) = x^2$.

The first one is known as L1-norm or Least Absolute Deviation (LAD) regression and it is introduced by Pierre–Simon Laplace (1749–1827) (Galileo Galilei (1564–1642) and Ruggiero Giovanni Boscovich (1711–1787)). The second one is known as Least Squares (LS) regression (it is the L2-norm) and it is introduced by Johann Carl Friedrich Gauss (1777–1855).

2.2.2 Ordinary Least Squares

Ordinary Least Squares (OLS) Principle

Let $\mathcal{B}$ be the set of all possible vectors $\boldsymbol{\beta}$. In general $\mathcal{B} = \mathbb{R}^k$. The object is to find a vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k) \in \mathcal{B}$ that minimize the sum of squared errors

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^t\boldsymbol{e} = (\boldsymbol{y} - X\boldsymbol{\beta})^t(\boldsymbol{y} - X\boldsymbol{\beta})$$

given the vector of observations $\boldsymbol{y}$ and the matrix of observations X.

- A minimum will always exists since $S(\boldsymbol{\beta})$ is a real–valued convex differentiable function.
- In the formulation of the OLS principle we do not need the normality assumption on Y but only that $\boldsymbol{y}$ is an independent sample from $Y \sim (E(Y|X = x), \sigma^2)$.

We may rewrite $S(\boldsymbol{\beta})$ as

$$\begin{aligned} S(\boldsymbol{\beta}) &= \boldsymbol{y}^t\boldsymbol{y} - \underbrace{\boldsymbol{\beta}^t X^t \boldsymbol{y}}_{1\times 1} - \underbrace{\boldsymbol{y}^t X \boldsymbol{\beta}}_{1\times 1} + \boldsymbol{\beta}^t X^t X \boldsymbol{\beta} \\ &= \boldsymbol{y}^t\boldsymbol{y} - 2\boldsymbol{\beta}^t X^t \boldsymbol{y} + \boldsymbol{\beta}^t X^t X \boldsymbol{\beta} \end{aligned}$$

and differentiate by $\boldsymbol{\beta}$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{\beta}) &= 2X^tX\boldsymbol{\beta} - 2X^t\boldsymbol{y} \\ \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^t} S(\boldsymbol{\beta}) &= 2X^tX \qquad \text{(non negative definite)} \end{aligned}$$

Equating the first derivative to zero yields what is called **normal equations**.

Normal Equations

$$X^tX\boldsymbol{\beta} = X^t\boldsymbol{y}$$

There are two cases:

1. X is of full rank k, then $\underbrace{X^tX}_{k\times k}$ is nonsingular and hence

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

Recall that $r\left((AA^t)\right) = r\left((A^tA)\right) = r\left(A\right) = r\left(A^t\right)$.

2. X is not of full rank then a set of solutions is given by

$$\hat{\boldsymbol{\beta}} = (X^t X)^- X^t \boldsymbol{y} + \left(I - (X^t X)^- (X^t X)\right) \boldsymbol{w}$$

where $(X^t X)^-$ is a g–inverse and $\boldsymbol{w}$ is an arbitrary vector of length k.

We will discuss, in details, the second case later on.

Definition 196 (Fitted values) Given a solution from the Ordinary Least Squares normal equations $\hat{\boldsymbol{\beta}}$ then

$$\begin{aligned} \hat{\boldsymbol{y}} &= X\hat{\boldsymbol{\beta}} \\ &= X(X^t X)^{-1} X^t \boldsymbol{y} && \text{if } X \text{ is of full rank} \\ &= H\boldsymbol{y} \end{aligned}$$

are the **fitted values**, that is, the empirical predictor of $\boldsymbol{y}$ by the linear model. △

Definition 197 (Residuals) The **residuals** $\boldsymbol{r}$ are defined as

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{y} - \hat{\boldsymbol{y}} \\ &= \boldsymbol{y} - X\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{y} - X(X^t X)^{-1} X^t \boldsymbol{y} && \text{if } X \text{ is of full rank} \\ &= (I - H)\boldsymbol{y} \\ &= \hat{\boldsymbol{e}} \ . \end{aligned}$$

By definition the residuals are an estimates of the (unobserved) errors $\boldsymbol{e}$. △

Theorem 198 H and $(I - H)$ are orthogonal projectors. □

Proof We have to prove that H is idempotent and symmetric.

$$H^2 = X\underbrace{(X^t X)^{-1} X^t X}_{I}(X^t X)^{-1} X^t = X(X^t X)^{-1} X^t = H$$

$$H^t = \left(X(X^t X)^{-1} X^t\right)^t = (X^t)^t (X^t X)^{-1^t} X^t = H \qquad \blacksquare$$

Theorem 199

- The fitted values $\hat{\boldsymbol{y}}$ have the same value for all solutions $\hat{\boldsymbol{\beta}}$ of the normal equations;
- $S(\boldsymbol{\beta})$, the sum of squares attains the minimum for any solution of the normal equations. □

Proof • Recall that $X(X^tX)^-X^tX = X$ then

$$X\hat{\boldsymbol{\beta}} = X(X^tX)^-X^t\boldsymbol{y} + \underbrace{X\left(I - (X^tX)^-(X^tX)\right)}_{\mathcal{O}}\boldsymbol{w}$$
$$= X(X^tX)^-X^t\boldsymbol{y}$$

which is independent of $\boldsymbol{w}$.

•

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\boldsymbol{y} - X\boldsymbol{\beta})^t(\boldsymbol{y} - X\boldsymbol{\beta}) \\ &= \left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}} + X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)^t \left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}} + X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) \\ &= (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) \\ &+ \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^t X^tX \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &+ 2\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^t X^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) \ . \end{aligned}$$

The last term is zero since

$$X^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = X^t\boldsymbol{y} - \underbrace{X^tX(X^tX)^-X^t}_{X^t}\boldsymbol{y}$$

Finally, since X^tX is nonnegative definite then

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^t X^tX \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \geq 0$$

hence

$$S(\boldsymbol{\beta}) \geq (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = S(\hat{\boldsymbol{\beta}})$$

and

$$S(\hat{\boldsymbol{\beta}}) = \boldsymbol{y}^t\boldsymbol{y} - 2\hat{\boldsymbol{\beta}}^tX^t\boldsymbol{y} + \hat{\boldsymbol{\beta}}^tX^tX\hat{\boldsymbol{\beta}}$$

by normal equations $\boldsymbol{y}^tX = (X^t\boldsymbol{y})^t = (X^tX\hat{\boldsymbol{\beta}})^t = \hat{\boldsymbol{\beta}}^tX^tX$

$$\begin{aligned} &= \boldsymbol{y}^t\boldsymbol{y} - \hat{\boldsymbol{\beta}}^tX^tX\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{y}^t\boldsymbol{y} - \hat{\boldsymbol{y}}^t\hat{\boldsymbol{y}} \\ &= \boldsymbol{r}^t\boldsymbol{r} \ . \end{aligned}$$

■

Geometric Properties of OLS

Let $\mathcal{R}(X) = \{\boldsymbol{\theta} : \boldsymbol{\theta} = X\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathbb{R}^k\}$ the column space generated by the column of the matrix X (the range space of the matrix X).

$\mathcal{R}(X)$ is a subspace of $\mathbb{R}^n$. If we equipped $\mathcal{R}(X)$ with the norm $\|\boldsymbol{x}\| = (\boldsymbol{x}^t\boldsymbol{x})^{1/2}$ for $\boldsymbol{x} \in \mathbb{R}^n$ then the OLS is the same as that of minimizing $\|\boldsymbol{y} - \boldsymbol{\theta}\|$ for $\boldsymbol{\theta} \in \mathcal{R}(X)$.

Theorem 200 The minimum of $\|\boldsymbol{y} - \boldsymbol{\theta}\|$ for $\boldsymbol{\theta} \in \mathcal{R}(X)$ is attained at $\hat{\boldsymbol{\theta}}$ such that $(\boldsymbol{y} - \hat{\boldsymbol{\theta}}) \perp \mathcal{R}(X)$ that is, when $(\boldsymbol{y} - \hat{\boldsymbol{\theta}})$ is orthogonal to all vectors in $\mathcal{R}(X)$, which is, when $\hat{\boldsymbol{\theta}}$ is the orthogonal projection of $\boldsymbol{y}$ on $\mathcal{R}(X)$. Such a $\hat{\boldsymbol{\theta}}$ exists and is unique, and has the explicit expression

$$\hat{\boldsymbol{\theta}} = H\boldsymbol{y} = X(X^tX)^{-1}X^t\boldsymbol{y} \ . \qquad \square$$

If X is not full rank then the theorem holds using a g–inverse.

Proof Let $\hat{\boldsymbol{\theta}} \in \mathcal{R}(X)$ be such that $(\boldsymbol{y} - \hat{\boldsymbol{\theta}}) \perp \mathcal{R}(X)$, that is, $X^t(\boldsymbol{y} - \hat{\boldsymbol{\theta}}) = \mathbf{0}$. Then

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{\theta}\|^2 &= \Big((\boldsymbol{y} - \hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\Big)^t \Big((\boldsymbol{y} - \hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\Big) \\ &= (\boldsymbol{y} - \hat{\boldsymbol{\theta}})^t(\boldsymbol{y} - \hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + 2(\boldsymbol{y} - \hat{\boldsymbol{\theta}})^t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\end{aligned}$$

but

$$(\boldsymbol{y} - \hat{\boldsymbol{\theta}})^t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \underbrace{(\boldsymbol{y} - \hat{\boldsymbol{\theta}})^t X}_{\mathbf{0}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

hence

$$\|\boldsymbol{y} - \boldsymbol{\theta}\|^2 = \|\boldsymbol{y} - \hat{\boldsymbol{\theta}}\|^2 + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \geq \|\boldsymbol{y} - \hat{\boldsymbol{\theta}}\|^2 \qquad \blacksquare$$

Notice that the orthogonality condition could be rewritten as (since $\hat{\boldsymbol{\theta}} = X\hat{\boldsymbol{\beta}}$)

$$X^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = \mathbf{0} \Leftrightarrow X^t\boldsymbol{y} = X^tX\hat{\boldsymbol{\beta}}$$

where on the left we have the normal equations provided by the OLS principle.

Bias of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{y}}$

Hereafter we are using the repeated sample principle to evaluate the bias and the variance of the introduced estimators and we are working conditioned to the values assumed by the matrix X. (Note that the normality assumption on the conditional distribution of $\boldsymbol{Y}$ is not needed).

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}((X^tX)^{-1}X^t\boldsymbol{Y}) \\ &= (X^tX)^{-1}X^t\,\mathrm{E}(\boldsymbol{Y}) \\ &= (X^tX)^{-1}X^tX\boldsymbol{\beta} \\ &= \boldsymbol{\beta}\end{aligned}$$

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{Y}}) &= \mathrm{E}(X\hat{\boldsymbol{\beta}}) \\ &= X\,\mathrm{E}(\hat{\boldsymbol{\beta}}) \\ &= X\boldsymbol{\beta}\ .\end{aligned}$$

Variance of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{Y}}$

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}}) &= (X^tX)^{-1}X^t\,\mathrm{var}(\boldsymbol{Y})\big((X^tX)^{-1}X^t\big)^t \\ &= (X^tX)^{-1}X^t(\sigma^2 I)X(X^tX)^{-1} \\ &= \sigma^2(X^tX)^{-1}\end{aligned}$$

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{Y}}) &= X\,\mathrm{var}(\hat{\boldsymbol{\beta}})X^t \\ &= \sigma^2X(X^tX)^{-1}X^t \\ &= \sigma^2H\ .\end{aligned}$$

Estimation of σ^2

An unbiased estimator of σ^2 is

$$\hat{\sigma^2} = \frac{\|\boldsymbol{r}\|^2}{n-k} = \frac{\sum_{i=1}^n r_i^2}{n-k} = \frac{\langle \boldsymbol{r},\, \boldsymbol{r}\rangle}{n-k}\ .$$

In fact

$$\begin{aligned}
\mathrm{E}(\|\boldsymbol{r}\|^2) &= \mathrm{E}\left(\boldsymbol{Y}^t(I-H)\boldsymbol{Y}\right) \\
&= (X\boldsymbol{\beta})^t\underbrace{(I-H)(X\boldsymbol{\beta})}_{\mathbf{0}} + \mathrm{E}\left(\boldsymbol{\varepsilon}^t(I-H)\boldsymbol{\varepsilon}\right) \\
&= \mathrm{E}\left(\mathrm{tr}\left(\boldsymbol{\varepsilon}^t(I-H)\boldsymbol{\varepsilon}\right)\right) \\
&= \mathrm{tr}\left((I-H)\,\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t)\right) \\
&= \mathrm{tr}\left((I-H)\sigma^2 I\right) \\
&= \sigma^2(n-k)
\end{aligned}$$

since the trace of idempotent matrix is equal to the rank and $r\,(I) = n$, $r\,(H) = k$.

Theorem 201 (Gauss–Markov) If $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 I)$ then the OLS estimator $\hat{\boldsymbol{\beta}}$ is such that

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) \leq \mathrm{var}((\tilde{\boldsymbol{\beta}}))$$

with respect to any other estimator $\tilde{\boldsymbol{\beta}}$ on the form $\tilde{\boldsymbol{\beta}} = C\boldsymbol{y}$, where C is a matrix of size $k \times n$ of constants such that $\mathrm{E}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. □

Since this theorem the OLS estimator $\hat{\boldsymbol{\beta}}$ is the Best Linear Unbiased Estimator (BLUE). Of course nonlinear estimators may have smaller variance.

Proof Since $\boldsymbol{\beta} = \mathrm{E}(\tilde{\boldsymbol{\beta}}) = \mathrm{E}(C\boldsymbol{Y}) = CX\boldsymbol{\beta}$ must holds for all $\boldsymbol{\beta}$ then $CX = I = X^tC^t$. Hence

$$\begin{aligned}
\underset{k\times k}{\mathrm{var}}(\tilde{\boldsymbol{\beta}}) - \underset{k\times k}{\mathrm{var}}(\hat{\boldsymbol{\beta}}) &= C\sigma^2 IC^t - \sigma^2(X^tX)^{-1} \\
&= \sigma^2\left(CIC^t - CX(X^tX)^{-1}X^tC^t\right) \\
&= \sigma^2 C(I-H)C^t \\
&= \sigma^2 C(I-H)(I-H)^tC^t\ .
\end{aligned}$$

and this $k \times k$ matrix is positive semidefinite. Thus $\hat{\boldsymbol{\beta}}$ has smallest variance in finite samples among all linear unbiased estimators of $\boldsymbol{\beta}$, provided that the second–order assumptions hold. ■

Example 202 (Straight–line regression model) We write the straight–line regression model in matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

The least squares estimates are

$$\begin{aligned}\boldsymbol{\beta} = \begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} &= \begin{bmatrix} n & \sum(x_i-\bar{x}) \\ \sum(x_i-\bar{x}) & \sum(x_i-\bar{x})^2\end{bmatrix}^{-1} \begin{bmatrix}\sum y_i \\ \sum(x_i-\bar{x})y_i\end{bmatrix} \\ &= \begin{bmatrix} n^{-1} & 0 \\ 0 & (\sum(x_i-\bar{x})^2)^{-1}\end{bmatrix} \begin{bmatrix}\sum y_i \\ \sum(x_i-\bar{x})y_i\end{bmatrix} \\ &= \begin{bmatrix}\bar{y} \\ \frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}\end{bmatrix}\end{aligned}$$

If all the x_i are equal, X^tX is not invertible, and $\hat{\beta}_1$ is indeterminated: any value is possible.

The unbiased estimator of σ^2 is

$$\hat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \bar{y} - (x_i-\bar{x})\frac{\sum_{j=1}^{n}(x_j-\bar{x})y_j}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right)^2$$

The variance of $\hat{\boldsymbol{\beta}}$ is

$$\operatorname{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \begin{bmatrix} n^{-1} & 0 \\ 0 & (\sum(x_i-\bar{x})^2)^{-1}\end{bmatrix} \qquad \boxtimes$$

2.2.3 Likelihood

Distribution of the errors

In a linear model we had assumed $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$ then

$$\begin{aligned} f(\boldsymbol{\varepsilon}) &= (2\pi)^{-n/2}\left|\sigma^2 I\right|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{\varepsilon}^t(\sigma^2 I)^{-1}\boldsymbol{\varepsilon}\right) \\ &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left(-\frac{1}{2\sigma^2}\varepsilon_i^2\right)\end{aligned}$$

Likelihood function

$$
\begin{aligned}
L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{e}) &= (2\pi)^{-n/2} \left|\sigma^2 I\right|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{e}^t(\sigma^2 I)^{-1}\boldsymbol{e}\right) \\
&= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2} e_i^2\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^t(\boldsymbol{y} - X\boldsymbol{\beta})\right) \\
&= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i{}^t\boldsymbol{\beta})^2\right)
\end{aligned}
$$

Log–likelihood function

$$
\begin{aligned}
\ell(\boldsymbol{\beta}, \sigma^2; \boldsymbol{e}) &= -\frac{1}{2}\left(n\log(2\pi) + n\log(\sigma^2) + \frac{1}{\sigma^2}\boldsymbol{e}^t\boldsymbol{e}\right) \\
&= -\frac{1}{2}\left(n\log(2\pi) + n\log(\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n} e_i^2\right) \\
&= -\frac{1}{2}\left(n\log(2\pi) + n\log(\sigma^2) + \frac{1}{\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^t(\boldsymbol{y} - X\boldsymbol{\beta})\right) \\
&= -\frac{1}{2}\left(n\log(2\pi) + n\log(\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i{}^t\boldsymbol{\beta})^2\right)
\end{aligned}
$$

Normal Equations

Whatever the value of σ^2, the log–likelihood is maximized with respect to $\boldsymbol{\beta}$ at the value that minimizes the sum of squares

$$
S(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^t(\boldsymbol{y} - X\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i{}^t\boldsymbol{\beta})^2 \ .
$$

We obtain the maximum likelihood estimate of $\boldsymbol{\beta}$ by solving simultaneously the equations

$$
\frac{\partial}{\partial\beta_r} = 2\sum_{i=1}^{n} x_{ir}(y_i - \boldsymbol{\beta}^t\boldsymbol{x}_i) = 0 \ , \qquad r = 1, \cdots, k \ .
$$

In matrix form these amount to

$$X^t(\boldsymbol{y} - X\boldsymbol{\beta}) = \mathbf{0}$$

which imply that the estimate satisfies $(X^tX)\boldsymbol{\beta} = X^t\boldsymbol{y}$.

Maximum Likelihood Estimator for $\boldsymbol{\beta}$

Provided the $k \times k$ matrix X^tX is of full rank, then

$$\hat{\boldsymbol{\beta}} = (X^tX)^{-1}X^t\boldsymbol{y}$$

that is the least squares estimator of $\boldsymbol{\beta}$.

Profile likelihood for σ^2

The maximum likelihood estimator of σ^2 may be obtained from the profile likelihood for σ^2

$$\ell_P(\sigma^2) = \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}\left(n\log(\sigma^2) + \frac{1}{\sigma^2}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})\right)$$

and it follows by differentiation that the maximum likelihood estimator of σ^2 is

$$\tilde{\sigma^2} = n^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = n^{-1}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i{}^t\hat{\boldsymbol{\beta}})^2 \ .$$

We shall see below that $\tilde{\sigma^2}$ is biased and that an unbiased estimator of σ^2 is

$$\hat{\sigma^2} = (n-k)^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = (n-k)^{-1}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i{}^t\hat{\boldsymbol{\beta}})^2 \ .$$

Fisher Information

The observed and expected information matrices play a central role in likelihood inference, by providing approximate variances for maximum likelihood estimates. To obtain these matrices for the normal linear model, note

that the log–likelihood has second derivatives

$$\frac{\partial^2}{\partial\beta_r\partial\beta_s}\ell(\boldsymbol{\beta},\sigma^2) = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ir}x_{is}$$
$$\frac{\partial^2}{\partial\beta_r\partial\sigma^2}\ell(\boldsymbol{\beta},\sigma^2) = \frac{1}{\sigma^4}\sum_{i=1}^{n} x_{ir}(y_i - \boldsymbol{x}_i{}^t\boldsymbol{\beta})$$
$$\frac{\partial^2}{\partial\sigma^2\partial\sigma^2}\ell(\boldsymbol{\beta},\sigma^2) = -\frac{1}{2}\left(-\frac{n}{\sigma^4} + \frac{2}{\sigma^6}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i{}^t\boldsymbol{\beta})^2\right)$$

for all $r, s = 1, \cdots, k$.

Thus elements of the expected information matrix are

$$\mathrm{E}\left(-\frac{\partial^2}{\partial\beta_r\partial\beta_s}\ell(\boldsymbol{\beta},\sigma^2)\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ir}x_{is}$$
$$\mathrm{E}\left(-\frac{\partial^2}{\partial\beta_r\partial\sigma^2}\ell(\boldsymbol{\beta},\sigma^2)\right) = 0$$
$$\mathrm{E}\left(-\frac{\partial^2}{\partial\sigma^2\partial\sigma^2}\ell(\boldsymbol{\beta},\sigma^2)\right) = \frac{n}{2\sigma^4}$$

or in matrix form

$$I(\boldsymbol{\beta},\sigma^2) = \begin{bmatrix}\sigma^{-2}X^tX & 0\\ 0 & \frac{1}{2}n\sigma^{-4}\end{bmatrix}, \quad I(\boldsymbol{\beta},\sigma^2)^{-1} = \begin{bmatrix}\sigma^2(X^tX)^{-1} & 0\\ 0 & 2\sigma^4/n\end{bmatrix}$$

Example 203

2.3 Inference using Likelihood and Bootstrap

Distribution of some quadratic forms

- $\boldsymbol{Y}^t\boldsymbol{Y} \sim \sigma^2\chi^2_n(\delta)$ with $\delta = \frac{1}{2}\boldsymbol{\beta}^tX^tX\boldsymbol{\beta}/\sigma^2$;
- $\boldsymbol{Y}^tH\boldsymbol{Y} \sim \sigma^2\chi^2_k(\delta)$ since twice the noncentrality parameter is

$$\begin{aligned}\boldsymbol{\beta}^tX^tHX\boldsymbol{\beta}/\sigma^2 &= \boldsymbol{\beta}^tX^t(X(X^tX)^{-1}X^t)X\boldsymbol{\beta}/\sigma^2\\ &= \boldsymbol{\beta}^t(\underbrace{X^tX(X^tX)^{-1}}_{I}X^tX)\boldsymbol{\beta}/\sigma^2\\ &= \boldsymbol{\beta}^tX^tX\boldsymbol{\beta}/\sigma^2\\ &= 2\delta\end{aligned}$$

- $\boldsymbol{Y}^t(I-H)\boldsymbol{Y} \sim \sigma^2\chi^2_{n-k}$ since the noncentrality parameter is

$$\begin{aligned}\boldsymbol{\beta}^t X^t(I-H)X\boldsymbol{\beta}/\sigma^2 &= \boldsymbol{\beta}^t X^t(I - X(X^tX)^{-1}X^t)X\boldsymbol{\beta}/\sigma^2 \\ &= \boldsymbol{\beta}^t(X^tX - \underbrace{X^tX(X^tX)^{-1}}_{I}X^tX)\boldsymbol{\beta}/\sigma^2 \\ &= 0\end{aligned}$$

System of Hypotheses

We first derive the Likelihood Ratio Test for the following system of hypotheses

$$\begin{cases} H_0 & : \boldsymbol{\beta} = \mathbf{0} \\ H_1 & : \boldsymbol{\beta} \neq \mathbf{0} \end{cases}$$

Later, we will consider a more general case and test based on Wald statistics.

2.3.1 Likelihood Ratio Test

Maximum Likelihood Estimators under the Null Hypothesis

Under the Null Hypothesis we have:

$$\hat{\boldsymbol{\beta}}_0 = \mathbf{0}$$

and the profile log–likelihood under the null hypothesis is

$$\ell_P(\sigma^2) = \ell(\mathbf{0}, \sigma^2) = -\frac{1}{2}\left(n\log(\sigma^2) + \frac{1}{\sigma^2}\boldsymbol{y}^t\boldsymbol{y}\right)$$

hence the score function is

$$\frac{\partial}{\partial\sigma^2}\ell_P(\sigma^2) = -\frac{1}{2}\left(\frac{n}{\sigma^2} - \frac{1}{\sigma^4}\boldsymbol{y}^t\boldsymbol{y}\right)$$

and equating it to zero leads to the solution

$$\tilde{\sigma}^2_0 = \frac{\boldsymbol{y}^t\boldsymbol{y}}{n}$$

Maximum Likelihood Estimators under the Alternative Hypothesis

As we already proved the MLE estimators without restrictions (that is under the Alternative Hypothesis) are

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}$$

$$\tilde{\sigma}_1^2 = \frac{(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})}{n}$$

Likelihood Ratio Test (LRT)

$$\begin{aligned}
\lambda(\boldsymbol{y}) &= \frac{\sup_{H_0} L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y})}{\sup_{H_1} L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y})} \\
&= \frac{(\tilde{\sigma}_0^2)^{-n/2} \exp\left(-\frac{1}{2}\tilde{\sigma}_0^{-2}\boldsymbol{y}^t\boldsymbol{y}\right)}{(\tilde{\sigma}_1^2)^{-n/2} \exp\left(-\frac{1}{2}\tilde{\sigma}_1^{-2}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_1)^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_1)\right)} \\
&= \frac{(\tilde{\sigma}_0^2)^{-n/2} \exp\left(-\frac{1}{2}n\right)}{(\tilde{\sigma}_1^2)^{-n/2} \exp\left(-\frac{1}{2}n\right)} \\
&= \left(\frac{(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_1)^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_1)}{\boldsymbol{y}^t\boldsymbol{y}}\right)^{n/2} \\
&= \left(\frac{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_1\|^2}{\|\boldsymbol{y}\|^2}\right)^{n/2}
\end{aligned}$$

Since

$$\begin{aligned}
\|\boldsymbol{y}\|^2 - \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_1\|^2 &= \boldsymbol{y}^t\boldsymbol{y} - ((I-H)\boldsymbol{y})^t((I-H)\boldsymbol{y}) \\
&= \boldsymbol{y}^t I\boldsymbol{y} - \boldsymbol{y}^t(I-H)\boldsymbol{y} \\
&= \boldsymbol{y}^t H\boldsymbol{y} \\
&= \|X\hat{\boldsymbol{\beta}}_1\|^2
\end{aligned}$$

and $\lambda(\boldsymbol{y})$ is equivalent to any monotone transformation, we have

$$\begin{aligned}
\lambda^*(\boldsymbol{y}) &= \lambda(\boldsymbol{y})^{-2/n} - 1 \\
&= \frac{\|X\hat{\boldsymbol{\beta}}_1\|^2}{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_1\|^2} \\
&= \frac{\|H\boldsymbol{y}\|^2}{\|(I-H)\boldsymbol{y}\|^2} \\
&= \frac{\boldsymbol{y}^t H\boldsymbol{y}}{\boldsymbol{y}^t(I-H)\boldsymbol{y}}
\end{aligned}$$

The distribution of $\lambda^*(\boldsymbol{Y})$ is given by the fact

$$F(\boldsymbol{Y}) = \lambda^*(\boldsymbol{Y})\frac{n-k}{k} = \frac{\frac{\boldsymbol{y}^t H \boldsymbol{y}}{k}}{\frac{\boldsymbol{y}^t (I-H) \boldsymbol{y}}{n-k}}$$

is an F of Snedecor with $k, n-k$ degrees of freedom and noncentral parameter equal to $\delta = \frac{1}{2}\boldsymbol{\beta}^t X^t X \boldsymbol{\beta}/\sigma^2$ which is 0 under the null hypothesis.

The Wilks (profile) log–likelihood ratio statistic is

$$\begin{aligned} -2\left(\ell_p(\tilde{\sigma}_0^2; \boldsymbol{Y}) - \ell_p(\tilde{\sigma}_1^2; \boldsymbol{Y})\right) &= n \log\left(\lambda^{-2/n}(\boldsymbol{Y})\right) \\ &= n \log(1 + \lambda^*(\boldsymbol{Y})) \\ &= n \log\left(1 + \frac{k}{n-k} F(\boldsymbol{Y})\right) \end{aligned}$$

and recall that, since the classical asymptotic theory this quantity is asymptotically distributed as a χ_k^2 under the null hypothesis.

Maximum Likelihood Estimator under (Linear) Restrictions

We are going to derive the MLE under linear restrictions on the values of the parameters $\boldsymbol{\beta}$, such restrictions can be specify using a matrix $R : q \times k$ (with $q \leq k$) and $r(R) = q$ so that

$$R\boldsymbol{\beta} = \mathbf{0}$$

These results would be very useful for solving the LRT statistic for a quite general null hypothesis.

The new function to minimized is

$$S_R(\boldsymbol{\beta}, \boldsymbol{\gamma}) = S(\boldsymbol{\beta}) + 2(R\boldsymbol{\beta})^t \boldsymbol{\gamma} = (\boldsymbol{y} - X\boldsymbol{\beta})^t (\boldsymbol{y} - X\boldsymbol{\beta}) + 2(R\boldsymbol{\beta})^t \boldsymbol{\gamma}$$

where $\boldsymbol{\gamma}$ is the vector of Lagrange multipliers. The derivative of $S_R(\boldsymbol{\beta}, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ leads to the following system of equations

$$\begin{cases} X^t X \boldsymbol{\beta} + R^t \boldsymbol{\gamma} = X^t \boldsymbol{y} \\ R\boldsymbol{\beta} = \mathbf{0} \end{cases}$$

The first equation can be rewritten as

$$\begin{aligned} \underbrace{(X^t X)^{-1} X^t X}_{I} \boldsymbol{\beta} + (X^t X)^{-1} R^t \boldsymbol{\gamma} &= (X^t X)^{-1} X^t \boldsymbol{y} \\ \boldsymbol{\beta} + (X^t X)^{-1} R^t \boldsymbol{\gamma} &= \hat{\boldsymbol{\beta}} \end{aligned}$$

that is

$$\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} - (X^t X)^{-1} R^t \boldsymbol{\gamma}$$

and using it on the second equation we have

$$\hat{\boldsymbol{\gamma}} = (R(X^t X)^{-1} R^t)^{-1} R\hat{\boldsymbol{\beta}}$$

that leads to the restricted estimate of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}} - (X^t X)^{-1} R^t \left(R(X^t X)^{-1} R^t\right)^{-1} R\hat{\boldsymbol{\beta}}$$

The restricted projection $\hat{\boldsymbol{y}}_R$ is

$$\begin{aligned}
\hat{\boldsymbol{y}}_R &= X\hat{\boldsymbol{\beta}}_R \\
&= X\left(\hat{\boldsymbol{\beta}} - (X^t X)^{-1} R^t \left(R(X^t X)^{-1} R^t\right)^{-1} R\hat{\boldsymbol{\beta}}\right) \\
&= X\hat{\boldsymbol{\beta}} - X(X^t X)^{-1} R^t \left(R(X^t X)^{-1} R^t\right)^{-1} R\hat{\boldsymbol{\beta}} \\
&= \hat{\boldsymbol{y}} - X(X^t X)^{-1} R^t \left(R(X^t X)^{-1} R^t\right)^{-1} R(X^t X)^{-1} X^t \boldsymbol{y} \\
&= H\boldsymbol{y} - H_R \boldsymbol{y} \\
&= (H - H_R)\boldsymbol{y}
\end{aligned}$$

where we let $H_R = X(X^t X)^{-1} R^t \left(R(X^t X)^{-1} R^t\right)^{-1} R(X^t X)^{-1} X^t$. H_R is an orthogonal projector.

The profile log–likelihood under linear restrictions is

$$\begin{aligned}
\ell_P(\sigma^2) &= \ell(\hat{\boldsymbol{\beta}}_R, \sigma^2) \\
&= -\frac{1}{2}\left(n\log(\sigma^2) + \frac{1}{\sigma^2}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_R)^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_R)\right) \\
&= -\frac{1}{2}\left(n\log(\sigma^2) + \frac{1}{\sigma^2}((I - H + H_R)\boldsymbol{y})^t((I - H + H_R)\boldsymbol{y})\right)
\end{aligned}$$

hence the score function is

$$\frac{\partial}{\partial\sigma^2}\ell_P(\sigma^2) = -\frac{1}{2}\left(\frac{n}{\sigma^2} - \frac{1}{\sigma^4}((I - H + H_R)\boldsymbol{y})^t((I - H + H_R)\boldsymbol{y})\right)$$

and equating it to zero leads to the solution

$$\tilde{\sigma}^2_R = \frac{((I - H + H_R)\boldsymbol{y})^t((I - H + H_R)\boldsymbol{y})}{n} = \frac{\boldsymbol{y}^t(I - H + H_R)\boldsymbol{y}}{n}\ .$$

Linear System of Hypotheses

We are going to consider the following system of hypotheses

$$\begin{cases} H_0 & : R\boldsymbol{\beta} = \mathbf{0} \\ H_1 & : R\boldsymbol{\beta} \neq \mathbf{0} \end{cases}$$

Maximum Likelihood Estimators under the Null Hypothesis

Under the Null Hypothesis we have:

$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_R$$

$$\tilde{\sigma}_0^2 = \tilde{\sigma}_R^2 = \frac{\boldsymbol{y}^t(I - H + H_R)\boldsymbol{y}}{n}$$

Likelihood Ratio Test (LRT)

$$\begin{aligned}
\lambda(\boldsymbol{y}) &= \frac{\sup_{H_0} L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y})}{\sup_{H_1} L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y})} \\
&= \frac{(\tilde{\sigma}_R^2)^{-n/2} \exp\left(-\frac{1}{2}\tilde{\sigma}_R^{-2}(\boldsymbol{y} - \hat{\boldsymbol{\beta}}_R)^t(\boldsymbol{y} - \hat{\boldsymbol{\beta}}_R)\right)}{(\tilde{\sigma}^2)^{-n/2} \exp\left(-\frac{1}{2}\tilde{\sigma}^{-2}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})\right)} \\
&= \frac{(\tilde{\sigma}_R^2)^{-n/2} \exp\left(-\frac{1}{2}n\right)}{(\tilde{\sigma}^2)^{-n/2} \exp\left(-\frac{1}{2}n\right)} \\
&= \left(\frac{(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})}{(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_R)^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_R)}\right)^{n/2} \\
&= \left(\frac{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2}{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_R\|^2}\right)^{n/2}
\end{aligned}$$

Since

$$\begin{aligned}
&= \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_R\|^2 - \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2 \\
&= ((I - H + H_R)\boldsymbol{y})^t(I - H + H_R)\boldsymbol{y} - ((I - H)\boldsymbol{y})^t((I - H)\boldsymbol{y}) \\
&= \boldsymbol{y}^t(I - H + H_R)\boldsymbol{y} - \boldsymbol{y}^t(I - H)\boldsymbol{y} \\
&= \boldsymbol{y}^t H_R \boldsymbol{y}
\end{aligned}$$

Recall that $\hat{\boldsymbol{y}}_R = (H - H_R)\boldsymbol{y}$ and hence $\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_R = H_R\boldsymbol{y}$

$$= \|X\hat{\boldsymbol{\beta}} - X\hat{\boldsymbol{\beta}}_R\|^2$$

and $\lambda(\boldsymbol{y})$ is equivalent to any monotone transformation, we have

$$\begin{aligned}
\lambda^*(\boldsymbol{y}) &= \lambda(\boldsymbol{y})^{-2/n} - 1 \\
&= \frac{\|X\hat{\boldsymbol{\beta}} - X\hat{\boldsymbol{\beta}}_R\|^2}{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2} \\
&= \frac{\|H_R\boldsymbol{y}\|^2}{\|(I-H)\boldsymbol{y}\|^2} \\
&= \frac{\boldsymbol{y}^t H_R \boldsymbol{y}}{\boldsymbol{y}^t (I-H)\boldsymbol{y}}
\end{aligned}$$

The distribution of $\lambda^*(\boldsymbol{Y})$ is given by the fact

$$F(\boldsymbol{Y}) = \lambda^*(\boldsymbol{Y})\frac{n-k}{q} = \frac{\frac{\boldsymbol{y}^t H_R \boldsymbol{y}}{q}}{\frac{\boldsymbol{y}^t (I-H)\boldsymbol{y}}{n-k}}$$

is an F of Snedecor with $q, n-k$ degrees of freedom and noncentral parameter equal to

$$\begin{aligned}
2\delta &= (\boldsymbol{\beta}^t X^t H_R X \boldsymbol{\beta})/\sigma^2 \\
&= \boldsymbol{\beta}^t X^t X (X^t X)^{-1} R^t \left(R(X^t X)^{-1} R^t\right)^{-1} R(X^t X)^{-1} X^t X \boldsymbol{\beta})/\sigma^2 \\
&= \boldsymbol{\beta}^t R^t \left(R(X^t X)^{-1} R^t\right)^{-1} R\boldsymbol{\beta})/\sigma^2
\end{aligned}$$

which is 0 under the null hypothesis.

The Wilks (profile) log–likelihood ratio statistic is

$$\begin{aligned}
-2\left(\ell_p(\tilde{\sigma}_0^2; \boldsymbol{Y}) - \ell_p(\tilde{\sigma}_1^2; \boldsymbol{Y})\right) &= n \log\left(\lambda^{-2/n}(\boldsymbol{Y})\right) \\
&= n \log(1 + \lambda^*(\boldsymbol{Y})) \\
&= n \log\left(1 + \frac{q}{n-k} F(\boldsymbol{Y})\right)
\end{aligned}$$

and recall that, since the classical asymptotic theory this quantity is asymptotically distributed as a χ^2_q under the null hypothesis.

Example 204 An important case is to test the value of only one component, say $1 \le r \le k$, that is

$$\begin{cases} H_0 & : \boldsymbol{\beta}_r = 0 \\ H_1 & : \boldsymbol{\beta}_r \neq 0 \end{cases}$$

For this system of hypotheses we can set

$$\underset{1\times k}{R} = (0, 0, \cdots, 1, \cdots, 0, 0)$$

where the one is in position r. Hence

$$\|X\hat{\boldsymbol{\beta}} - X\hat{\boldsymbol{\beta}}_R\|^2 = \|H_R\boldsymbol{y}\|^2 = \hat{\boldsymbol{\beta}}_r^2/(X^tX)^{-1}{}_{r,r} = \hat{\beta}_r^2/v_{rr}$$

where v_{rr} is the element in position (r, r) of the matrix $V = (X^tX)^{-1}$. The F statistics is

$$F(\boldsymbol{y}) = \frac{\hat{\boldsymbol{\beta}}_r^2}{\hat{\sigma^2}v_{rr}}$$

and $F(\boldsymbol{Y})$ has distribution F of Snedecor with $(1, n-k)$ degrees of freedom. Recall that $\hat{\sigma^2}$ is the unbiased estimator of σ^2. The square root of this test statistic

$$t(\boldsymbol{y}) = \sqrt{F(\boldsymbol{y})} = \frac{\hat{\boldsymbol{\beta}}_r}{\hat{\sigma}\sqrt{v_{rr}}}$$

is the well known t statistics where $t(\boldsymbol{Y})$ has a t of student distribution with $n-k$ degrees of freedom. ⊠

2.3.2 Confidence Intervals for components of the parameter vector

Probabilistic based confidence intervals

The usual way of constructing (univariate) confidence intervals is to use the fact

$$\hat{\boldsymbol{\beta}}_i \sim \mathrm{N}(\boldsymbol{\beta}_i, \sigma^2 v_{ii})$$

where v_{ii} is the ith diagonal element of $(X^tX)^{-1}$, and $\hat{\boldsymbol{\beta}}$ is independent of $\hat{\sigma^2}$ (Show why!), whose distribution is $(n-k)^{-1}\sigma^2\chi^2_{n-k}$. Therefore

$$T = \frac{\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i}{\sqrt{\hat{\sigma^2}v_{ii}}} \sim t_{n-k} \ .$$

A $(1-\alpha)$ confidence interval for $\boldsymbol{\beta}_i$ is

$$\hat{\boldsymbol{\beta}}_i \pm \hat{\sigma} v_{ii}^{1/2} t_{n-k,1-\frac{\alpha}{2}} \ .$$

Confidence Ellipsoids for the Whole Parameter Vector $\boldsymbol{\beta}$

To construct a confidence ellipsoids for the whole parameter vector $\boldsymbol{\beta}$ we can use the results on the likelihood ratio test

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t X^t X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(\boldsymbol{y}-X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y}-X\hat{\boldsymbol{\beta}})}\frac{n-k}{k} = \frac{\|X\hat{\boldsymbol{\beta}}-X\boldsymbol{\beta}\|^2}{\|\boldsymbol{y}-X\hat{\boldsymbol{\beta}}\|^2}\frac{n-k}{k} \leq F_{k,n-k,1-\alpha}$$

Likelihood based confidence intervals

A Likelihood based confidence intervals for the whole parameter vector $\boldsymbol{\beta}$ can be constructed using the normalized likelihood function

$$\frac{L(\boldsymbol{\beta},\sigma^2;\boldsymbol{y})}{\sup_{H_1} L(\boldsymbol{\beta},\sigma^2;\boldsymbol{y})} \leq \lambda$$

where λ is a suitable constant.

3 Analysis of Variance

3.1 Analysis of variance approach to regression analysis

Analysis of variance approach to (simple) regression analysis

Analysis of variance is based on the decomposition of the Total Sum of Square (SST) in Regression Sum of Square (SSR) and Error Sum of Square (SSE)

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SSE}$$

The Mean Square (MS) is obtained dividing the Sum of Squares by its degrees of freedom.

ANOVA Table

Source of Variation	SS	df	MS	F
Regression	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	1	$MSR=\frac{SSR}{1}$	$\frac{MSR}{MSE}$
Error	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$n-2$	$MSE=\frac{SSE}{n-2}$	
Total	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$		

Analysis of Variance provides a F–test for testing:

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$$

- Test statistic:

$$F_{obs} = \frac{MSR}{MSE}$$

- By Cochran's Theorem, we have for the corresponding statistics

$$F \overset{H_0}{\sim} F_{(1,n-2)}$$

- Decision rule:
  - If $F_{obs} > F_{(1-\alpha;1,n-2)}$ reject H_0;
  - If $F_{obs} \leq F_{(1-\alpha;1,n-2)}$ accept H_0.
- Equivalence of F–test and t–test in simple linear regression.

General Linear Test Approach

Three basic steps:

1. Fit the full model and obtain the error sum of squares $SSE(F)$
2. Fit the reduced model under H_0 and obtain the error sum of squares $SSE(R)$
3. Use the F statistic

$$F_{obs} = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$$

which follows the F distribution under H_0 (df_R and df_F are the degrees of freedom associated with $SSE(R)$ and $SSE(F)$ respectively).

- Decision rule:
  - If $F_{obs} > F_{(1-\alpha;df_R-df_F,df_F)}$ reject H_0;
  - If $F_{obs} \leq F_{(1-\alpha;df_R-df-F,df_F)}$ accept H_0.

Example 205 In simple linear regression for testing whether or not $\beta_1 = 0$ we have:

$$SSE(R) = SST \quad SSE(F) = SSE$$
$$df_R = n-1 \qquad df_F = n-2$$

then,

$$F_{obs} = \frac{\frac{SST-SSE}{(n-1)-(n-2)}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE} \qquad \boxtimes$$

Descriptive measures of association between X and Y

- Coefficient of Determination ($0 \leq R^2 \leq 1$)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

  R^2 measures the proportionate reduction of total variation associated with the use of the predictor variable X;

- Coefficient of Correlation, in the simple regression model we have

$$\rho_{X,Y} = \pm\sqrt{R^2} \ .$$

F test of lack of fit

Objective: to test if a specific type of regression is adequate for the data

$$\begin{cases} H_0: & E(Y) = \beta_0 + \beta_1 X \\ H_1: & E(Y) \neq \beta_0 + \beta_1 X \end{cases}$$

- Fit the full model and obtain the error sum of squares $SSE(F)$;
- Fit the reduced model under H0 and obtain the error sum of squares $SSE(R)$;
- Use the F statistic

$$F_{obs} = \frac{\frac{SSE(R)-SSE(F)}{df_R-df_F}}{\frac{SSE(F)}{df_F}}$$

  which under H_0 follows an F distribution.

F test of lack of fit with replicates

- Setup
  - Y are independent and normally distributed with constant variance;
  - Repeat observations at one or more X levels (replicates)
- Notation
  - J number of different levels of X, i.e. $x_1, \cdots, x_J$;
  - n_j number of replicates for the j–th level of X;
  - Y_{ij} value of the i–th replicate at level j–th of X;
  - $i = 1, \cdots, n_j$ and $j = 1, \cdots, J$;
  - $\bar{y}_j$ mean of Y observations at $X = x_j$.
- Full Model

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \qquad \text{where } \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$SSE(F) = \sum_{j}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

$$df_F = n - J$$

- Reduced Model

$$Y_{ij} = \beta_0 + \beta_1 x_j + \varepsilon_{ij}, \qquad \text{where } \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$SSE(R) = \sum_{j}^{J} \sum_{i=1}^{n_j} (y_{ij} - \hat{y}_{ij})^2$$

$$df_R = n - 2$$

- Test Statistic

$$F_{obs} = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$$

where $SSE(F) = SSPE$, $SSE(R) = SSE$,

$$F_{obs} = \frac{\frac{SSE-SSPE}{(n-2)-(n-J)}}{\frac{SSPE}{n-J}}$$

- Defining the lack of fit sum of squares as $SSLF = SSE - SSPE$

$$F_{obs} = \frac{\frac{SSLF}{(J-2)}}{\frac{SSPE}{n-J}} = \frac{MSLF}{MSPE}$$

- Under H_0

$$F \sim F_{J-2,n-J}$$

- Decision rule:
  - If $F_{obs} > F_{(1-\alpha;J-2,n-J)}$ reject H_0;
  - If $F_{obs} \leq F_{(1-\alpha;J-2,n-J)}$ accept H_0.
- Anova Table for Lack of Fit

Source of Variation	SS	df	MS
Regression (SSR)	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\hat{y}_{ij} - \bar{y})^2$	1	$MSR = \frac{SSR}{1}$
Error (SSE)	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \hat{y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit (SSLF)	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\bar{y}_{ij} - \hat{y}_{ij})^2$	$J-2$	$MSLF = \frac{SSLF}{J-2}$
Pure error (SSPE)	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{ij})^2$	$n-J$	$MSPE = \frac{SSPE}{n-J}$
Total (SST)	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2$	$n-1$	

3.2 Analysis of variance in linear models

Analysis of variance in a (general) linear models

- Total Sum of Squares

$$\begin{aligned} SST &= \boldsymbol{y}^t\boldsymbol{y} - \frac{1}{n}\boldsymbol{y}^t J\boldsymbol{y} \\ &= \boldsymbol{y}^t(I - \frac{1}{n}J)\boldsymbol{y} \end{aligned}$$

where J is a matrix of 1's of dimension $n \times n$;

- Residual Sum of Squares

$$\begin{aligned} SSE = \boldsymbol{r}^t\boldsymbol{r} &= (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}^t\boldsymbol{y} - \hat{\boldsymbol{\beta}}^t X^t\boldsymbol{y} \\ &= \boldsymbol{y}^t(I - H)\boldsymbol{y} \end{aligned}$$

- Regression (Model) Sum of Squares

$$SSR = \hat{\boldsymbol{\beta}}^t X^t\boldsymbol{y} - \frac{1}{n}\boldsymbol{y}^t J\boldsymbol{y} = \boldsymbol{y}^t\left(H - \frac{1}{n}J\right)\boldsymbol{y}$$

ANOVA Table for the General Linear Regression

- ANOVA Table

Source of Variation	SS	df	MS
Regression	SSR	$p-1$	$MSR = \frac{SSR}{p-1}$
Error	SSE	$n-p$	$MSE = \frac{SSE}{n-p}$
Total	SST	$n-1$	

- Analysis of Variance provides a F-test for testing:

$$\begin{cases} H_0: & \boldsymbol{\beta} = \mathbf{0} \\ H_1: & \boldsymbol{\beta} \neq \mathbf{0} \end{cases}$$

- Statistic

$$F_{obs} = \frac{MSR}{MSE}$$

- Decision rule:
  - If $F_{obs} > F_{(1-\alpha;p-1,n-p)}$ reject H_0;
  - If $F_{obs} \leq F_{(1-\alpha;p-1,n-p)}$ accept H_0.
- Coefficient of determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Adjust coefficient of determination

$$R_a^2 = 1 - \frac{MSE}{MST} = 1 - \frac{n-1}{n-p}\frac{SSE}{SST}$$

Extra Sum of Squares

Marginal increase of the regression sum of squares when one or several predictor variables are added in the regression model/marginal reduction of the error sum of squares when one or several predictor variables are added in the model.

Example 206 The extra sum of squares when adding variable X_2 in the simple regression model with X_1 is

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

or

$$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1) \ .$$

Then $SSR(X_2|X_1)$ is the reduction in the error sum of square when adding X_2 in a regression model with X_1.

Decomposition of SSR into Extra Sum of Squares

$$SSR(X_1, X_2) = SSR(X_2) + SSR(X_1|X_2) = SSR(X_1) + SSR(X_2|X_1) \quad \boxtimes$$

Example 207 With three or more variables

$$\begin{aligned} SSR(X_3|X_1, X_2) &= SSE(X_1, X_2) - SSE(X_1, X_2, X_3) \\ &= SSR(X_1, X_2, X_3) - SSR(X_1, X_2) \end{aligned}$$

and

$$\begin{aligned} SSR(X_2, X_3|X_1) &= SSE(X_1) - SSE(X_1, X_2, X_3) \\ &= SSR(X_1, X_2, X_3) - SSR(X_1) \end{aligned}$$

Decomposition of SSR into Extra Sum of Squares

$$\begin{aligned} SSR(X_1, X_2, X_3) &= SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) \\ &= SSR(X_2) + SSR(X_3|X_2) + SSR(X_1|X_2, X_3) \\ &= SSR(X_3) + SSR(X_2|X_3) + SSR(X_1|X_2, X_3) \end{aligned}$$

and all others possible permutations of the three indices. $\boxtimes$

ANOVA Table with Extra Sum of Squares

Source of variation	SS	df	MS
Regression	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2 \mid X_1$	$SSR(X_2 \mid X_1)$	1	$MSR(X_2 \mid X_1)$
$X_3 \mid X_1, X_2$	$SSR(X_3 \mid X_1, X_2)$	1	$MSR(X_3 \mid X_1, X_2)$
Error	SSE	$n-4$	$MSE(X_1, X_2, X_3)$
Total	SST	$n-1$	$MST(X_1, X_2, X_3)$

This is one of the possible ANOVA Table we can construct by permutation of the three indices.

Test for Regression Coefficients (using Extra Sum of Squares)

- Test whether a single $\beta_k = 0$

$$\begin{cases} H_0: & \beta_k = 0 \\ H_1: & \beta_k \neq 0 \end{cases}$$

in a linear model with p explanatory variables. To test use the general linear test approach.

- Full Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_F$

$$SSE(F) = SSE(X_1, \cdots, X_p)$$

$$df_F = n - p$$

- Reduced Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_1 X_{i(k-1)} + \beta_1 X_{i(k+1)} + \cdots + \beta_p X_{ip} + \varepsilon_R$

$$SSE(F) = SSE(X_1, \cdots, X_{k-1}, X_{k+1}, \cdots X_p)$$

$$df_R = n - p + 1$$

- The general linear test statistic is

$$F_{obs} = \frac{\frac{SSE(X_1,\cdots,X_{k-1},X_{k+1},\cdots X_p) - SSE(X_1,\cdots,X_p)}{(n-p+1)-(n-p)}}{\frac{SSE(X_1,\cdots,X_p)}{n-p}}$$

$$= \frac{SSR(X_k \mid X_1, \cdots, X_{k-1}, X_{k+1}, \cdots X_p)}{MSE(X_1, \cdots, X_p)} = \frac{MSR(X_k \mid X_1, \cdots, X_{k-1}, X_{k+1}, \cdots X_p)}{MSE(X_1, \cdots, X_p)}$$ p

- Test whether several $\beta_k = 0$

$$\begin{cases} H_0: & \beta_k = \beta_j = 0 \\ H_1: & \text{not both } \beta_k \text{ and } \beta_j \text{ are equal to } 0 \end{cases}$$

- Full Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_F$

$$SSE(F) = SSE(X_1, \cdots, X_p)$$

$$df_F = n - p$$

- Reduced Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_1 X_{i(k-1)} + \beta_1 X_{i(k+1)} + \cdots + \beta_1 X_{i(j-1)} + \beta_1 X_{i(j+1)} + \cdots + \beta_p X_{ip} + \varepsilon_R$

$$SSE(F) = SSE(X_1, \cdots, X_{-k}, \cdots, X_{-j}, \cdots X_p)$$

$$df_R = n - p + 2$$

- The general linear test statistic is

$$\begin{aligned} F_{obs} &= \frac{\frac{SSE(X_1,\cdots,X_{-k},\cdots,X_{-j},\cdots X_p) - SSE(X_1,\cdots,X_p)}{(n-p+2)-(n-p)}}{\frac{SSE(X_1,\cdots,X_p)}{n-p}} \\ &= \frac{SSR(X_k, X_j | X_1, \cdots, X_{-k}, \cdots, X_{-j}, \cdots, X_p)/2}{MSE(X_1, \cdots, X_p)} \\ &= \frac{MSR(X_k, X_j | X_1, \cdots, X_{-k}, \cdots, X_{-j}, \cdots, X_p)}{MSE(X_1, \cdots, X_p)} \end{aligned}$$

  partial F test statistic.

- Other type of test than

$$\begin{cases} H_0: & \beta_k = \beta_j \\ H_1: & \beta_k \neq \beta_j \end{cases}$$

or

$$\begin{cases} H_0: & \beta_k = 3, \beta_j = 5 \\ H_1: & \beta_k \neq 3, \beta_j \neq 5 \end{cases}$$

cannot be done using Extra Sum of Squares. In these cases you have to use the Likelihood Ratio Test Statistic.

Coefficients of partial determination

Consider $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon$ (Coefficient of partial determination of Y and X_1, given that X_2 is in the model: relative marginal reduction in the variation of Y associated with X_1 when X_2 is already in the model)

$$R^2_{Y1.2} = \frac{SSR(X_1|X_2)}{SSE(X_2)}$$

The coefficient of partial determination when there are 3 or more variables in the model

$$R^2_{Y1.23} = \frac{SSR(X_1|X_2, X_3)}{SSE(X_2, X_3)}$$

$$R^2_{Y4.123} = \frac{SSR(X_4|X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)}$$

The coefficients of partial correlation are the square root of the coefficients of partial determination.

4 Model Checking

4.1 Introduction

4.1.1 Specification

4.1.2 Checking

Model Checking

Discrepancies between data and a linear model may be isolated or systematic, or both.

- One type of isolated discrepancy is when there are outliers: a few observations that are unusual relative to the rest.
- Systematic discrepancies arise, for example, when a transformation of the response or a covariate is needed, when correlated errors are supposed independent, or when a term is incorrectly omitted.

There are many techniques for detecting such problems. Graphs are widely used, often supplemented by more formal methods.

Assumptions in the linear model

- **lineratity**: the response depends linearly on each explanatory variable and on the error, with no systematic dependence on any omitted terms;

- **constant variance** (homoschedasticity): the responses have equal variances, which in particular do not depend on the level of the response;
- **independence**: the errors are uncorrelated, and independent if normal; and sometimes
- **normality**: in the normal linear model the errors are normally distributed.

Prediction (or Hat) Matrix

The matrix

$$H = X(X^tX)^{-1}X^t$$

is called the **prediction** (or **hat**) matrix. It is symmetric and idempotent of $r(P) = \mathrm{tr}(P) = \mathrm{tr}(I_k) = k$.

The matrix

$$M = I - H$$

is calles the **residuals** matrix. It is also symmetric and idempotent with $r(M) = n - k$.

Understanding the properties of these matrices is very important for the model checking process.

The (i, j)th element of the matrix H is denoted by h_{ij} where

$$h_{ij} = h_{ji} = \boldsymbol{x}_j{}^t(X^tX)^{-1}\boldsymbol{x}_i \qquad i, j = 1, \cdots, n \ .$$

The ex-post predictor $\hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}} = H\boldsymbol{Y}$ has the dispersion matrix

$$\mathrm{var}(\hat{\boldsymbol{Y}}) = \sigma^2 H \ .$$

Therefore, we obtain

$$\begin{aligned} \mathrm{var}(\hat{Y}_i) &= \sigma^2 h_{ii} \\ \mathrm{var}(\boldsymbol{r}) &= \mathrm{var}((I-H)\boldsymbol{Y}) = \sigma^2(I-H) = \sigma^2 M \\ \mathrm{var}(r_i) &= \sigma^2(1-h_{ii}) \end{aligned}$$

and for $i \neq j$

$$\mathrm{cov}(r_i, r_j) = -\sigma^2 h_{ij} \ .$$

The correlation coefficient between r_i and r_j then becomes

$$\text{cor}(r_i, r_j) = \frac{-h_{ij}}{\sqrt{1-h_{ii}}\sqrt{1-h_{jj}}} \ .$$

Thus the covariance matrices of the predictor $\hat{\boldsymbol{Y}}$ and the estimator of error are entirely determined by H.

Observe that

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i}^{n} h_{ij} y_j \qquad 1 \leq i \leq n \ ,$$

implying that

$$\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii} \quad \text{and} \quad \frac{\partial \hat{y}_i}{\partial y_j} = h_{ij}$$

- h_{ii} can be interpreted as the amount of *leverage* each value y_i has in determining $\hat{y}_i$ regardless of the realized value y_i;
- h_{ij} can be interpreted as the influence of y_j in determining $\hat{y}_i$.

Using the model

$$\boldsymbol{y} = \mathbf{1}\alpha + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

we obtain

$$H = \frac{\mathbf{1}\mathbf{1}^t}{n} + \bar{X}(\bar{X}^t\bar{X})^{-1}\bar{X}^t$$

and

$$h_{ii} = \frac{1}{n} + \bar{x}_i^t(\bar{X}^t\bar{X})^{-1}\bar{x}_i$$

where $\bar{X}_{ij} = x_{ij} - \bar{x}_i$ is the matrix of the mean corrected x–values.

Because of symmetry of H, we have $h_{ij} = h_{ji}$, and the idempotence of H implies

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \ .$$

From this equation we obtain the important property

$$0 \leq h_{ii} \leq 1 \ .$$

We have also

$$h_{ii} = h_{ii}^2 + h_{ij}^2 + \sum_{k \neq i,j} h_{ik}^2 \qquad (j \text{ fixed}) \ .$$

which implies that $h_{ij}^2 \leq h_{ii}(1 - h_{ii})$, and therefore, we obtain

$$-0.5 \leq h_{ij} \leq 0.5 \qquad (i \neq j) \ .$$

- If $h_{ii} = 1$ or $h_{ii} = 0$ then $h_{ij} = 0$; that is, if a diagonal element h_{ii} is close to either 1 or 0, then the elements h_{ij} (for all $j \neq i$) are close to 0. For instance, the first component of the predictor $\hat{\boldsymbol{y}}$ is such that $\hat{y}_1 = \sum_{j=1}^{n} h_{1j} y_j$. If h_{11} is close to 0, then y_1 itself and all the other observations $y_2, \cdots, y_n$ have low influence on $\hat{y}_1$.;
- $(h_{ii}h_{jj} - h_{ij}^2) \geq 0$;
- $(1 - h_{ii})(1 - h_{jj}) - h_{ij}^2 \geq 0$;
- $h_{ii} + \frac{r_i^2}{\boldsymbol{r}^t\boldsymbol{r}} \leq 1$. If h_{ii} is large, then the standardized residual $r_i^2/(\boldsymbol{r}^t\boldsymbol{r})$ becomes small.

5 Model Selection

5.1 Introduction

What is the Best Model to Use?

While hundreds of books and countless journal papers deal with estimation of model parameters and their associated precision, relatively little has appeared concerning

- model specification: what set of candidate models to consider;
- model selection: what model(s) to use for inference;
- how to make inference after we choose a model(s).

In fact, Fisher believed at one time that model specification was outside the field of mathematical statistics, and this attitude prevailed within statistical community until at least the early 1970s.

What is the Best Model to Use?

- If one has data and a model, Likelihood theory can be used to estimate the unknown parameters (θ) and other quantities useful in making inference.

- However, which model is the best to use for making inferences? What is the basis for saying a model is "best"?
- If a poor or inappropriate model is used, then inference based on the data and this model will often be poor.
- Thus, it is clearly important to select (i.e. infer) an appropriate model for the analysis of a specific data set; however, this is not the same as trying to find the "true model".

Example 208 Flather [1992] and Flather [1996] studied patterns of avian species–accumulation rates among forested landscapes in the eastern United States using index data form the Breeding Bird Survey [Bystrak, 1981]. He derived an a priori set of 9 candidate model from two source:

1. the literature on species area curves (most often the power or exponential models were suggested);
2. a broader search of the literature for functions that increased monotonically to an asymptote. ⊠

Which model should be used for the analysis of these ecological datas?

Model structure	Number of parameters
$E(y) = ax^b$	3
$E(y) = a + b\log(x)$	3
$E(y) = a(x/(b+x))$	3
$E(y) = a(1 - \exp(-bx))$	3
$E(y) = a - bc^x$	4
$E(y) = (a + bx)/(1 + cx)$	4
$E(y) = a(1 - \exp -bx)^c$	4
$E(y) = a\left(1 - [1 + (x/c)^d]^{-b}\right)$	5
$E(y) = a[1 - 1/\exp\left(b(x-c)\right)^d]$	

Model building process

- Motivation;
- Data collection and preparation;
- Model specification;
- Model selection;
- Inference on the parameters;
- Model Validation.

Motivation

What will be the use of the model(s)?

- Understanding reality (causation);
- Approximation;
- Prediction;
- Forecast;

5.2 Model Specification

Model Specification

- Model specification (or formulation), in its widest sense, is conceptually more difficult than estimating the model parameters and their precision;
- Model specification is the point where the scientific information formally enter the investigation;
- Building the set of candidate models is partially a subjective art: that is why scientists must be trained, educated, and experienced in their discipline;
- The published literature and experience in the field can be used to help formulate a set of a priori candidate models;
- Good approximating models, each representing a scientific hypothesis, in conjunction with a good set of relevant data can provide insight into the underlying process and structure.

Model Specification and Explanatory Data Analysis

Exploring your data

- Flexible attitude;
- plots the data (using different methods);
- avoid computation of test statistics, P–values, and so forth.

Tukey concludes that to implement the confirmatory paradigm properly we need to do a lot of explanatory work.

Model Specification and a Global Model

- Development of the a priori set of candidate models often should include a global (or Full) model: a model that has many parameters, includes all potentially relevant effects, and reflects causal mechanisms thought likely, based on the "science of the situation";
- The global model should also reflect the study design and attributes of the system studied.

Model Specification and a Global Model

- The more parameters used, the better the fit of the model to the data that is achieved. Large and extensive data sets are likely to support more complexity, and this should be considered in the development of the set of candidate models.
- In developing the set of candidate models, one must recognize a certain balance between keeping the set small and focused on plausible hypotheses, while making it big enough to guard against omitting a very good a priori model.
- If a particular model (parameterization) does not make sense in the field of interest, this is reason to exclude it from the set of candidate models, particularly in the case where **causation** is of interest.

Models Versus Full Reality

- None of the models considered as the basis for data analysis are the "true model" that generates the data we observe;
- The "truth" in the sciences has essentially infinite dimension, and hence full reality cannot be revealed with only finite samples of data and a "model" of those data.
- A model is a simplification or approximation of reality and hence will not reflect all of reality.

J. Box: "all models are wrong, but some are useful".

The Principle of Parsimony

- William of Occam suggested in the fourteenth century that one "shave away all the is unnecessary" a dictum often referred to as *Occam's razor*. Occam's razor has had a long history in both science and technology, and it is embodied in the principle of parsimony.
- Albert Einstein: "Everything should be made as simple as possible, but not simpler".
- Statisticians view the principle of parsimony as a bias versus variance tradeoff. In general, bias decreases and variance increases as the dimension (complexity) of the model increases.

5.3 Model Selection

Model selection bias and uncertainty

When data are used to both select a parsimonious model and estimate the model parameters and their precision there are two sources of distortion

- Model selection bias;
- Model selection uncertainty.

Model selection bias

Example 209 Consider a linear model with

- a response variable (y);
- 4 explanatory variables x_j ($j = 1, \cdots, 4$).
- For convenience let x_1 very important, x_2 important, x_3 somewhat important, while x_4 is barely important.
- Given a decent sample size, nearly any model selection method will indicate that x_1 and probably x_2 are important ("dominant").
- If one had 1000 replicate data sets of the same size, from the same stochastic process
  - x_1 (particularly) and x_2 would be included in the model in nearly all cases;

- For models that included x_1 and x_2 (essentially all) the estimators of the coefficients would have good statistical properties w.r.t. bias and precision;
- For x_3 assume $|\beta_3|/se(\beta_3) \approx 1$, then this variable might be included in the model in only $15 - 30\%$ of the 1000 data sets.
  - in the model with x_3 the estimated regression coefficient is biased away from zero, inference in this case tend to exaggerate the importance of x_3;
  - in the model without x_3 imply that x_3 was of no importance.
  - neither of these cases is satisfactory.
- For x_4 assume $|\beta_4|/se(\beta_4) \approx 1/4$, this variable might be included in only a few $5 - 10\%$ of the data sets.
  - in the model with x_4 the estimated regression coefficient is large biased away from zero, inference in this case x_4 would be consider of great importance and the t–statistic $\hat{\beta}_4/\hat{se}(\hat{\beta}_4)$ is significant, because the numerator is biased high, while the denominator is biased low;
- when x_3 and x_4 are included in models, the associated estimator for a σ^2 is negatively biased and precision exaggerated. ⊠

Model selection bias

These two type of biases:

- of the estimators of the parameters (positively bias);
- of the estimators of the residual variance (negatively bias);

are called **model selection bias** and they can be quite serious.

Model selection bias

- When sample size is large, true replication exists, and there are relatively few models, model selection bias may be relatively unimportant.
- When one has only a small sample size, no true replication, and many models and variables; then model selection bias is usually severe.

Model selection uncertainty

The sampling variance of an estimator $\hat{\theta}$ has to components:

- var($\hat{\theta}$|model) which is the usual sampling variance
- a variance component due to not knowing the best approximating model to use (and, therefore, having to estimate this).

Failure to allow for model selection uncertainty often results in estimated sampling variance and covariances that are too low, and thus the achieved confidence interval coverage will be below the nominal value.

Optimal methods for coping with model selection uncertainty are at the forefront of statistical research.

Model selection uncertainty is problematic in masking statistical inferences.

5.4 Generating all subsets

Model selection approach

Two different approaches for choosing inside the candidate models:

- Consider all possible subsets (useful for sets of explanatory variables that are small or moderate in size, says not more than 20 variables);
- Consists in using automatic search procedures to arrive to a single subset (not all possible submodels are compared): e.g. Stepwise regression (recommended for reductions involving large pools of explanatory variables).

Generating all subsets

The obvious disadvantage of generating all subsets of a given Global Model is cost (and time).

If we have p variables than we have:

$$2^p - 1 = \sum_{i=1}^{p} \binom{p}{i}$$

submodels. Thus, the computational cost roughly doubles with each additional variable.

- $p = 2$ then $2^p - 1 = 3$
- $p = 4$ then $2^p - 1 = 15$

- $p = 8$ then $2^p - 1 = 255$
- $p = 16$ then $2^p - 1 = 65535$
- $p = 20$ then $2^p - 1 = 1048575$

5.5 Akaike Information Criterion

Notation

- $\boldsymbol{y}, \boldsymbol{x}$ are two independent samples from r.v.s Y and X both with distribution F and density f;
- $\mathcal{M}(k) = \{m(x; \boldsymbol{\theta}_k); \boldsymbol{\theta}_k \in \Theta(k)\}$ is a parametric family indexed by a k–dimensional parameters;
- $\mathcal{M} = \{\mathcal{M}_1, \cdots, \mathcal{M}_p\}$ is the set of all possible model (after model specification step);
- The goal is to search the collection of classes $\mathcal{M}$ for the "best" approximation to f.

Introduction

- We use concepts from Information theory, which has been a discipline only since the mid–1940s and covers a variety of theories and methods that are fundamental to many of the sciences [Cover and Thomas, 1991];
- The Kullback–Leibler "distance" or "information", between two models [Kullback and Leibler, 1951] is the main tool;
- Akaike [1973] found a simple relation between Kullback–Leibler distance and Fisher's maximized log–likelihood function;
- This relationship leads to a simple, effective, and very general methodology for selecting a parsimonious model for the analysis of empirical data.

Kullback and Leibler

Dr. Solomon Kullback
1903–1994

Dr. Richard A. Leibler
1914–2003

Definition 210 (Kullback–Leibler Information) Kullback–Leibler information between the two models f and m is defined as

$$I(f, m(x;\boldsymbol{\theta})) := \int f(x) \log\left(\frac{f(x)}{m(x;\boldsymbol{\theta})}\right) \, d\nu(x) \ ,$$

where log denotes the natural logarithm. △

- f is the full reality; $m(x;\boldsymbol{\theta})$ is a model;
- $I(f, m(x;\boldsymbol{\theta}))$ is a disparity between f and $m(x;\boldsymbol{\theta})$;
- $I(f, m(x;\boldsymbol{\theta}))$ could be interpret as the "information lost when $m(x;\boldsymbol{\theta})$ is used to approximate f";
- Is the negative of Boltzmann's generalized entropy [Boltzmann, 1877] in physics and thermodynamics;
- Shannon [1948] employed entropy in his famous treatise on communication theory.

Truth f drops out as a constant

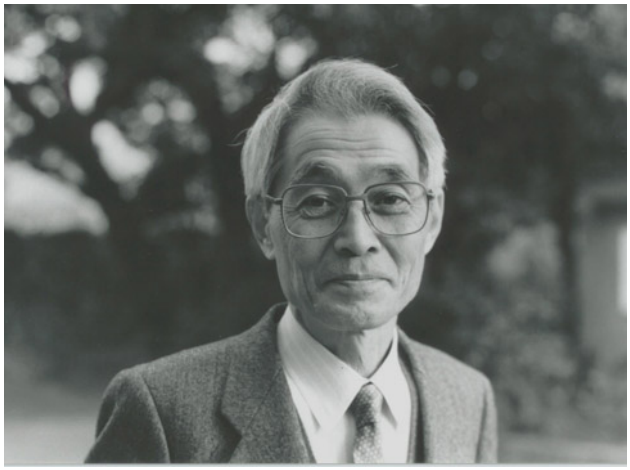
$$I(f, m(x;\boldsymbol{\theta})) = \int f(x) \log f(x)\ d\nu(x) - \int f(x) \log m(x;\boldsymbol{\theta})\ d\nu(x)$$

hence, each factor is a statistical expectation with respect to f

$$= \mathrm{E}_X\left(\log f(x)\right) - \mathrm{E}_X\left(\log m(x;\boldsymbol{\theta})\right)\ .$$

- The first expectation is a constant that depends only on the unknown true distribution, and it is clearly not known;
- Treating this unknown term as a constant leads to a **relative** measure between f and $m(x;\boldsymbol{\theta})$.
- Kullback–Leibler disparity is on a ratio scale, in fact there is a natural zero, while the second term is on an interval scale and lacks a natural zero.

Akaike



Dr. Hirotugu Akaike, born November 5, 1927. In 2006 Akaike was awarded the Kyoto Prize for his major contribution to statistical science and modeling in the development of the "Akaike Information Criterion" (AIC).

Akaike idea

The critical issue for getting an applied K–L model selection criterion was to estimate

$$E_Y E_X\left(\log m(x;\hat{\boldsymbol{\theta}}(y))\right)$$

where X and Y are independent random variables with the same distribution f and both statistical expectations are taken with respect to f. This expression is the target of all model selection approaches, based on K–L information and it is the relevant part of the **Expected Kullback–Leibler divergence**.

Akaike [1973] showed that the maximized log–likelihood

$$\ell(\hat{\boldsymbol{\theta}}(x); x)$$

is biased upward as an estimator of the model selection target. The bias is linked to the complexity of the model. Under certain conditions this bias is approximately equal to the number of estimable parameters k in the approximating model. That is

$$\ell(\hat{\boldsymbol{\theta}}(x); x) - k = \text{ constant } - \hat{\mathrm{E}}_{\hat{\boldsymbol{\theta}}} \left(I(f, m(\hat{\boldsymbol{\theta}})) \right)$$

and for large samples and "good" models

$$\ell(\hat{\boldsymbol{\theta}}(x); x) - k \approx E_Y E_X \left(\log \left(m(x; \hat{\boldsymbol{\theta}}(y)) \right) \right) \ .$$

Akaike's Information Criterion

Definition 211 (AIC) The Akaike's Information Criterion is

$$AIC := -2\ell(\hat{\boldsymbol{\theta}}(x); x) + 2k \qquad \triangle$$

- The AIC is an estimate of the expected, relative disparity between the unknown true mechanism that actually generated the observed data and the fitted model;
- The first term tends to decrease as more parameters are added to the approximating model, the second term gets larger as more parameters are added to the approximating model;
- This is a tradeoff between bias and variance or a tradeoff between underfitting and overfitting that is fundamental to the principle of parsimony.

Further comments

- Usually, AIC is positive; however it can be shifted by any additive constant, and some shifts can result in negative values of AIC;
- It is not the absolute size of the AIC value, it is the relative values over the set of models considered, and particularly the differences between AIC values, that are important.

Derivation of AIC

Main assumptions in the derivation of AIC are

- There exists a $k \leq p$ such that $f \in \mathcal{M}_k$;
- We denote by $\boldsymbol{\theta}_0$ the parameter vector such that $f(x) = m(x; \boldsymbol{\theta}_0)$ (a.e.);
- Conditions under which the consistency and asymptotic normality of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_k$ hold.

We will show that

$$\begin{aligned} -2E_Y E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(y)) \right) &= -2E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x)) \right) + 2k + o(1) \\ &= \mathrm{E}_X \left(AIC \right) + o(1) \end{aligned}$$

$$\begin{aligned} &= -2E_Y E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(y)) \right) \\ &= -2E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x)) \right) \\ &- 2 \left[E_X \left(\log m(x; \boldsymbol{\theta}_0) \right) - E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x)) \right) \right] \qquad (3) \\ &- 2 \left[E_Y E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(y)) \right) - E_X \left(\log m(x; \boldsymbol{\theta}_0) \right) \right] \qquad (4) \end{aligned}$$

The following Lemma asserts that (3) and (4) are both within $o(1)$ of k.

Lemma 212

$$\begin{aligned} -2 \left[E_X \left(\log m(x; \boldsymbol{\theta}_0) - E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x)) \right) \right] &= k + o(1) \\ -2 \left[E_Y E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(y)) \right) - E_X \left(\log m(x; \boldsymbol{\theta}_0) \right) \right] &= k + o(1) \qquad \square \end{aligned}$$

Proof We have to recall that

- Observed Fisher Information

$$\mathcal{I}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta}_k \partial {\boldsymbol{\theta}_k}^t} \ell(\boldsymbol{\theta}_k; \boldsymbol{x}) \ ;$$

- Expected Fisher Information

$$I(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}_0} \left[-\frac{\partial^2}{\partial \boldsymbol{\theta}_k \partial {\boldsymbol{\theta}_k}^t} \ell(\boldsymbol{\theta}_k; \boldsymbol{x}) \right] \ ;$$

First, consider taking a second–order Taylor expansion of $-2 \log m(x; \boldsymbol{\theta}_0)$ about $\hat{\boldsymbol{\theta}}_k(x)$, and evaluating the expectation of the result

$$\begin{aligned} -2 \log m(x; \boldsymbol{\theta}_0) &= -2 \log m(x; \hat{\boldsymbol{\theta}}_k(x)) \\ &\quad + \left(\hat{\boldsymbol{\theta}}_k(x) - \boldsymbol{\theta}_0 \right)^t \mathcal{I}(\hat{\boldsymbol{\theta}}(x)) \left(\hat{\boldsymbol{\theta}}_k(x) - \boldsymbol{\theta}_0 \right) \\ &\quad + o(1) \end{aligned}$$

Thus,

$$\begin{aligned} &= -2 \, \mathrm{E}_X \left(\log m(x; \boldsymbol{\theta}_0) - \log m(x; \hat{\boldsymbol{\theta}}_k(x)) \right) \\ &= \mathrm{E}_X \left(\left(\hat{\boldsymbol{\theta}}_k(x) - \boldsymbol{\theta}_0 \right)^t \mathcal{I}(\hat{\boldsymbol{\theta}}_k(x)) \left(\hat{\boldsymbol{\theta}}_k(x) - \boldsymbol{\theta}_0 \right) \right) + o(1) \end{aligned}$$

Next, consider taking a second–order Taylor expansion of $-2 E_Y \log m(x; \hat{\boldsymbol{\theta}}(y))$ about $\boldsymbol{\theta}_0$, and evaluating the expectation of the result

$$\begin{aligned} -2 E_Y E_X \log m(x; \hat{\boldsymbol{\theta}}_k(y)) &= -2 E_X \log m(x; \boldsymbol{\theta}_0) \\ &\quad + E_Y \left(\left(\hat{\boldsymbol{\theta}}_k(y) - \boldsymbol{\theta}_0 \right)^t I(\boldsymbol{\theta}_0) \left(\hat{\boldsymbol{\theta}}_k(y) - \boldsymbol{\theta}_0 \right) \right) \\ &\quad + o(1) \end{aligned}$$

Thus,

$$\begin{aligned} &= -2 \, \mathrm{E}_Y \, E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(y)) - \log m(x; \boldsymbol{\theta}_0) \right) \\ &= \mathrm{E}_Y \left(\left(\hat{\boldsymbol{\theta}}_k(y) - \boldsymbol{\theta}_0 \right)^t I(\boldsymbol{\theta}_0) \left(\hat{\boldsymbol{\theta}}_k(y) - \boldsymbol{\theta}_0 \right) \right) + o(1) \end{aligned}$$

The quadratic forms

$$\begin{aligned} &\left(\hat{\boldsymbol{\theta}}_k(x) - \boldsymbol{\theta}_0 \right)^t \mathcal{I}(\hat{\boldsymbol{\theta}}_k(x)) \left(\hat{\boldsymbol{\theta}}_k(x) - \boldsymbol{\theta}_0 \right) \\ &\left(\hat{\boldsymbol{\theta}}_k(y) - \boldsymbol{\theta}_0 \right)^t I(\boldsymbol{\theta}_0) \left(\hat{\boldsymbol{\theta}}_k(y) - \boldsymbol{\theta}_0 \right) \end{aligned}$$

both converge to centrally distributed chi–square random variable with k degrees of freedom, since we are assuming that $\boldsymbol{\theta}_0 \in \Theta(k)$, thus the expectations of both quadratic forms are within $o(1)$ of k. ■

Final considerations

- AIC provides us with an approximately unbiased estimator of the (relative part of the) Expected Kullback–Leibler divergence;
- The fitted model is either *correctly specified* or *overfitted*;
- AIC can still be effectively used in setting where $\mathcal{M}$ includes underspecified models;
- In setting where n is small and k is comparatively large (e.g. $k \approx n/2$, $2k$ is often much smaller than the bias adjustment, making AIC substantially negatively biased;
- If AIC severely underestimates the (relative part of the) Expected Kullback–Leibler divergence for high dimensional fitted models in the candidate set, the criterion may favor the higher dimensional models even when the expected discrepancy between these models and the generating model is rather large.
- In small–sample situations there exist better estimators of the bias adjustment term.

Example 213 (Linear models) If all the models in the set assume normally distributed errors with constant variance, then AIC can be easily computed from least squares as

$$AIC = n \log \hat{\sigma}^2 + 2k \ ,$$

where $\hat{\sigma}^2$ is the maximum likelihood estimator of the error variance which is

$$\hat{\sigma}^2 = \frac{\boldsymbol{e}^t \boldsymbol{e}}{n}$$

- The $\hat{\sigma}^2$ is not the corrected (unbiased) estimator of the error variance;
- Here k is the total number of parameters, including the intercept and σ^2. ⊠

Example 214 (ARMA models)

$$AIC(p, q) = -2\ell(\hat{\boldsymbol{\vartheta}}(x); x) + 2(p + q + 1) \qquad \boxtimes$$

AIC and log–likelihood ratio test

Let us consider the following hypotheses set

$$\begin{cases} H_0: & \beta_i = 0 \qquad \forall (q+1) < i \leq p \\ H_1: & \beta_i \neq 0 \qquad \forall 1 < i \leq p \end{cases}$$

for a set of nested models in which the full model is parameterized by $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ and σ^2 parameters. Let $\hat{\beta}_{H_0}$ be the maximum likelihood estimator under the null hypothesis and similarly for $\hat{\beta}_{H_1}$. Then the likelihood ratio test statistics is

$$\lambda = -2\left(\ell(\hat{\beta}_{H_0}; \hat{\sigma}^2) - \ell(\hat{\beta}_{H_1}; \hat{\sigma}^2)\right)$$

and, by definition of AIC we have

$$\lambda = 2(p-q) - (AIC_{H_1} - AIC_{H_0}) \ .$$

5.6 Takeuchi's Information Criterion

Bias correction when the model is not inside $\mathcal{M}$

- Pseudo true parameter

$$\bar{\boldsymbol{\theta}}_k = \text{argmin}_{\boldsymbol{\theta}_k \in \Theta(k)} \, I(f, m(\boldsymbol{\theta}_k)) \ ;$$

- when $f = m(\boldsymbol{\theta}_0) \in \mathcal{M}_k$ then $\bar{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_0$;
- when $f = m(\boldsymbol{\theta}_0) \notin \mathcal{M}_k$ then $\bar{\boldsymbol{\theta}}_k$ is such that $m(\bar{\boldsymbol{\theta}}_k)$ provides the best approximation to f (in the sense of Kullback-Leibler information).

Definition 215 (Takeuchi's Information Criterion)

$$TIC := -2\ell(\hat{\boldsymbol{\theta}}_k; \boldsymbol{x}) + 2\hat{\text{tr}}\big(J(\bar{\boldsymbol{\theta}}_k) I(\bar{\boldsymbol{\theta}}_k)^{-1}\big)$$

where

$$J(\boldsymbol{\theta}_k) = \text{E}_{\boldsymbol{\theta}_0}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}_k}\ell(\boldsymbol{\theta}_k; \boldsymbol{x})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}_k}\ell(\boldsymbol{\theta}_k; \boldsymbol{x})\right)^t\right]$$

is the variance of the score function and the last term is an estimate of

$$\text{tr}\left(J(\bar{\boldsymbol{\theta}}_k) I(\bar{\boldsymbol{\theta}}_k)^{-1}\right) \qquad \triangle$$

- In general $J(\bar{\boldsymbol{\theta}}_k)$ and $I(\bar{\boldsymbol{\theta}}_k)$ are unknown, since
  - $\bar{\boldsymbol{\theta}}_k$ is unknown and it is, in general, estimates by $\hat{\boldsymbol{\theta}}_k$;
  - $J(\cdot)$ and $I(\cdot)$ depend on θ_0;
- If θ_0 is estimate by $\hat{\boldsymbol{\theta}}_k$ the penalty term of TIC will often reduce to $2k$;
- Often, the Observed Fisher Information evaluated in $\hat{\boldsymbol{\theta}}_k$ is used to estimate $I(\hat{\boldsymbol{\theta}}_k)$ and similarly,

$$\left(\frac{\partial}{\partial \boldsymbol{\theta}_k}\ell(\hat{\boldsymbol{\theta}}_k)\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}_k}\ell(\hat{\boldsymbol{\theta}}_k)\right)^t$$

  is used to estimate $J(\hat{\boldsymbol{\theta}}_k)$.

Some computational issues and comparison

- Shibata [1989] notes that estimation error of the two matrices J and I can causes instability of the results on model selection. Consider the case where the candidate model has $k = 20$ parameters. Then J and I have both $k(k+1)/2 = 210$ values to be estimates. A reliable estimation will be difficult unless sample size is very large;
- $\text{tr}\,(JI^{-1})$ itself has a very simple parsimonious estimator, namely k;
- Thus AIC is an approximation to TIC;
- The approximation is excellent when the approximating model is "good" and becomes poor when the approximating model is poor. However, for models that are poor, the term related to log–likelihood dominates the criterion because the fit is poor;
- While TIC is an important contribution to the literature, it has rarely seen application. We suggest its use only when sample size is very large.

Example 216 (Linear models) For normal linear regression setting we have

$$TIC = -2\ell(\hat{\boldsymbol{\theta}}(x); x) + 2\left[\frac{\hat{\sigma_p^2}}{\hat{\sigma}_k^2}\left(k + 2 - \frac{\hat{\sigma_p^2}}{\hat{\sigma}_k^2}\right)\right]$$

- where k denotes the rank of the design matrix for the candidate model of interest;
- σ_p^2 denotes the maximum likelihood estimator of the error variance associated with the largest candidate model. ⊠

Derivation of TIC

$$
\begin{aligned}
&= -2E_Y E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(y))\right) \\
&= -2E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x))\right) \\
&- 2\left[E_X \left(\log m(x; \bar{\boldsymbol{\theta}}_k)\right) - E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x))\right)\right] \\
&- 2\left[E_Y E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(y))\right) - E_X \left(\log m(x; \bar{\boldsymbol{\theta}}_k)\right)\right] \\
&= -2E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x))\right) \\
&+ \mathrm{E}_X \left(\left(\hat{\boldsymbol{\theta}}_k(x) - \bar{\boldsymbol{\theta}}_k\right)^t \mathcal{I}(\hat{\boldsymbol{\theta}}_k(x)) \left(\hat{\boldsymbol{\theta}}_k(x) - \bar{\boldsymbol{\theta}}_k\right)\right) \\
&+ \mathrm{E}_Y \left(\left(\hat{\boldsymbol{\theta}}_k(y) - \bar{\boldsymbol{\theta}}_k\right)^t I(\bar{\boldsymbol{\theta}}_k) \left(\hat{\boldsymbol{\theta}}_k(y) - \bar{\boldsymbol{\theta}}_k\right)\right) \\
&+ o(1) \\
&= -2E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x))\right) \\
&+ 2\,\mathrm{E}_Y \left(\left(\hat{\boldsymbol{\theta}}_k(y) - \bar{\boldsymbol{\theta}}_k\right)^t I(\bar{\boldsymbol{\theta}}_k) \left(\hat{\boldsymbol{\theta}}_k(y) - \bar{\boldsymbol{\theta}}_k\right)\right) + o(1) \\
&= -2E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x))\right) \\
&+ 2\,\mathrm{tr}\left\{I(\bar{\boldsymbol{\theta}}_k)\,\mathrm{E}_Y \left(\left(\hat{\boldsymbol{\theta}}_k(y) - \bar{\boldsymbol{\theta}}_k\right)^t \left(\hat{\boldsymbol{\theta}}_k(y) - \bar{\boldsymbol{\theta}}_k\right)\right)\right\} + o(1) \\
&= -2E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x))\right) \\
&+ 2\,\mathrm{tr}\left\{I(\bar{\boldsymbol{\theta}}_k)\Sigma(\bar{\boldsymbol{\theta}}_k)\right\} + o(1) \\
&= -2E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x))\right) \\
&+ 2\,\mathrm{tr}\left\{I(\bar{\boldsymbol{\theta}}_k)I(\bar{\boldsymbol{\theta}}_k)^{-1}J(\bar{\boldsymbol{\theta}}_k)I(\bar{\boldsymbol{\theta}}_k)^{-1}\right\} + o(1) \\
&= -2E_X \left(\log m(x; \hat{\boldsymbol{\theta}}_k(x))\right) \\
&+ 2\,\mathrm{tr}\left\{J(\bar{\boldsymbol{\theta}}_k)I(\bar{\boldsymbol{\theta}}_k)^{-1}\right\} + o(1)
\end{aligned}
$$

where $\Sigma(\bar{\boldsymbol{\theta}}_k)$ is the large–sample variance and covariance matrix of $\hat{\boldsymbol{\theta}}_k$ which has the sandwich form

$$\Sigma(\bar{\boldsymbol{\theta}}_k) := I(\bar{\boldsymbol{\theta}}_k)^{-1}J(\bar{\boldsymbol{\theta}}_k)I(\bar{\boldsymbol{\theta}}_k)^{-1} \ .$$

5.7 A Small Sample Size AIC

Second Order Akaike Information Criterion

Definition 217 Hurvich and Tsai [1989] studied small–sample (second order) bias adjustment, which led to a criterion that is called AIC_c

$$\begin{aligned} AIC_c &:= -2\ell(\hat{\boldsymbol{\theta}}(x); x) + 2k\frac{n}{n-k-1} \\ &= -2\ell(\hat{\boldsymbol{\theta}}(x); x) + 2k + \frac{2k(k+1)}{n-k-1} \\ &= AIC + \frac{2k(k+1)}{n-k-1} \ . \end{aligned}$$ △

- Unless the sample size is large with respect to the number of estimated parameters, use of AIC_c is recommended;
- If n is large with respect to k (and p), then the second–order correction is negligible and AIC should perform well.

5.8 Modified Akaike Information Criterion

Modified Akaike Information Criterion

Definition 218

$$\begin{aligned} MAIC &:= -2\ell(\hat{\boldsymbol{\theta}}(x); x) + \frac{2n(k+1)}{n-k-2} \\ &\quad + \left[2k\left(\frac{(n-k)\hat{\sigma}_p^2}{(n-p)\hat{\sigma}_k^2} - 1 \right) - 2\left(\frac{(n-k)\hat{\sigma}_p^2}{(n-p)\hat{\sigma}_k^2} - 1 \right)^2 \right] \end{aligned}$$

where p denote the rank of the design matrix for the largest candidate model. △

Summary

	$f \in \mathcal{M}(k)$	$f \notin \mathcal{M}(k)$
Large samples	AIC	TIC
Small samples	AIC_c	MAIC

5.9 Schwarz Information Criterion

Schwarz Information Criterion

Definition 219

$$SIC := -2\ell(\hat{\boldsymbol{\theta}}(x); x) + k \ln n \ . \qquad \triangle$$

- AIC and SIC feature the same goodness–of–fit term;
- The penalty term of SIC is more stringent than the penalty term of AIC (for $n \geq 8$, $k \ln n \geq 2k$);
- SIC tends to favor smaller models than AIC.

Overview of SIC

- The Schwarz information criterion is often called the Bayesian information criterion;
- Common acronyms: SIC, BIC, SBC, SC;
- AIC provides an asymptotically unbiased estimator of the expected Kullback discrepancy between the generating model and the fitted approximating model;
- SIC provides a large-sample estimator of a transformation of the Bayesian posterior probability associated with the approximating model;
- By choosing the fitted candidate model corresponding to the minimum value of SIC, one is attempting to select the candidate model corresponding to the highest Bayesian posterior probability.

Overview of SIC

- SIC was justified by Schwarz [1978] "for the case of independent, identically distributed observations, and linear models", under the assumption that the likelihood is from the regular exponential family.
- Generalizations of Schwarz's derivation are presented by Stone [1979], Leonard [1982], Haughton [1988], and Cavanaugh and Neath [1999]

Generalized Information Criterion

Definition 220 (GIC)

$$GIC := -2\ell(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + a_n k \ .$$

where a_n represents a sequence that depends on the sample size n and possibly the dimension k. △

Method	a_n	Comments
AIC	2	
AIC_c	$2n/n - k - 2$	$\lim_{n\to\infty} a_n = 2$
SIC	$\log(n)$	

5.10 Bootstrap variants of AIC

Introduction

- In the normal univariate regression and multivariate regression frameworks, under the assumption $f \in \mathcal{M}(k)$, AICc is exactly unbiased for the expected Kullback–Leibler discrepancy;
- In all other frameworks for which AICc has been justified, under the assumption $f \in \mathcal{M}(k)$, AICc is only approximately unbiased for the expected Kullback–Leibler discrepancy;
- AICc has not been justified for many important modeling frameworks, including ordinary generalized linear models, mixed models, and state-space models (in time series analysis).

Bootstrap variants of AIC

- Bootstrapping has been used to formulate AIC variants where the bias adjustment is estimated via the bootstrap;
- In general, bootstrapping provides a means for assessing bias and variability;
- Recall that the goodness-of-fit term $-2\ell(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y})$ provides a (negatively) biased estimate of the expected Kullback–Leibler discrepancy $E_Y E_X \left(\log m(x; \hat{\boldsymbol{\theta}}(y))\right)$;
- With bootstrap-corrected variants of AIC, the bootstrap is used to adjust $-2\ell(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y})$ for this bias.

Bootstrap variants of AIC

- Let $\{\hat{\boldsymbol{\theta}}_k^*(i) : i = 1, \cdots, B\}$ represent a collection of B bootstrap replicates of $\hat{\boldsymbol{\theta}}_k$ corresponding to the B bootstrap samples.

Bootstrap variants of AIC: WIC

- The idea of using the bootstrap to improve the performance of a model selection rule was introduced by Efron [1983, 1986a]. Ishiguro and Sakamoto [1991] advocated a bootstrap variant of AIC, namely WIC, which is based on Efron's methodology. WIC is defined as

$$WIC := -2\ell(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + \left(\frac{1}{B} \sum_{i=1}^{B} -2 \log \frac{m(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_k^*(i))}{m(\boldsymbol{y}^*(i); \hat{\boldsymbol{\theta}}_k^*(i))} \right) .$$

Bootstrap variants of AIC

- Cavanaugh and Shumway [1997] proposed a bootstrap variant of AIC, AICb, for state-space model selection;
- Shibata [1997] has established the asymptotic equivalence of AICb and WIC under a general set of conditions, and has indicated the existence of other asymptotically equivalent bootstrap AIC variants.
- AICb is defined as

$$AICb := -2\ell(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + \left(\frac{1}{B} \sum_{i=1}^{B} -2 \log \frac{m(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_k^*(i))}{m(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_k)} \right) .$$

Bootstrap variants of AIC

- Advantages of bootstrap-corrected AIC variants:
  - Justifications do not require the restrictive assumption $f \in \mathcal{M}(k)$;
  - Justifications often require only a general set of conditions;
  - The criteria perform well in small to moderate sample-size applications;
- Disadvantages of bootstrap AIC variants:
  - The criteria are less straightforward to evaluate than AIC or AICc;
  - Evaluation may be computationally expensive.

s

6 Bootstrap

6.1 Introduction

Bootstrap: Introduction

What is the bootstrap? It is a general technique for estimating unknown quantities associated with statistical models. Often the bootstrap is used to find

- estimate the distribution of a statistics;
- standard errors for estimators;
- confidence intervals for unknown parameters;
- p values for test statistics under a null hypothesis.

Thus the bootstrap is typically used to estimate quantities associated with the sampling distribution of estimators and test statistics.

The bootstrap was introduced in 1979 as a computer–based (intensive) method for estimating the standard error of an estimator [Efron, 1979].

Definition 221 (Empirical Distribution Function) Having observed a random sample $\boldsymbol{x} = (x_1, \cdots, x_n)$ of size n from an unknown distribution F the **empirical distribution function** $\hat{F}_n$ is defined to be the discrete distribution that puts probability $1/n$ on each value x_i, $1 \leq i \leq n$. In other words, $\hat{F}_n$ assigns to a set A in the sample space of X its empirical probability

$$\widehat{\Pr(A)} = \frac{\#(x_i \in A)}{\#\boldsymbol{x}}$$

the proportion of the observed sample $\boldsymbol{x}$ occuring in A. The formula is

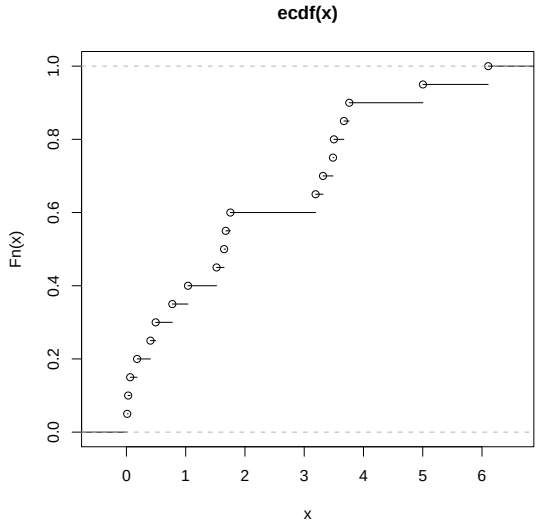
$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{(-\infty,t]}(x_i) \qquad \triangle$$

Example 222 We consider the following sample of size $n = 20$:

0.013	0.029	0.063	0.180	0.405	0.494	0.774	1.039	1.521	1.648
1.676	1.754	3.192	3.317	3.485	3.501	3.670	3.760	5.004	6.105

$\boxtimes$

The corresponding empirical distribution function is



Definition 223 (Plug–in Principle) The **plug–in principle** is a simple method of estimating parameters from samples. The **plug–in estimate** of a parameter $\theta = t(F)$ is defined to be

$$\hat{\theta} = t(\hat{F}_n) \ .$$

In other words, we estimate the function $\theta = t(F)$ of the probability distribution F by the same function of the empirical distribution $\hat{F}_n$, i.e., $\hat{\theta} = t(\hat{F}_n)$.△

Example 224 Using the dataset presented in the example 222. If we are interesting in $\theta = \int x^2 dF(x)$ which is the second moment of the r.v. with distribution F then, by the plug–in principle, an estimate of θ is

$$\hat{\theta} = \int x^2 d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 7.428 \ .$$

Here we are not assuming anything about the shape of F. This situation is called the **nonparametric setup**. ⊠

The Bootstrap

We will illustrate the main idea of bootstrap by the simple example of estimating a standard deviation of the estimate $\hat{\theta}$ of a parameter $\theta = t(F)$. We have a random sample $\boldsymbol{x} = (x_1, \cdots, x_n)$ of size n from an unknown distribution F and let $\hat{F}$ be the empirical distribution function.

A **bootstrap sample** is defined to be a random sample of size n drawn from $\hat{F}$, say

$$\boldsymbol{x}^* = (x_1^*, x_2^*, \cdots, x_n^*) \ .$$

The star notation indicates that $\boldsymbol{x}^*$ is not the actual data set $\boldsymbol{x}$, but rather a randomized, or *resampled*, version of $\boldsymbol{x}$.

A bootstrap sample can be viewed as data points $x_1^*, x_2^*, \cdots, x_n^*$ that are a random sample of size n drawn with replacement from the population of n objects $(x_1, \cdots, x_n)$.

Corresponding to a bootstrap sample $\boldsymbol{x}^*$ is a **bootstrap replication** of $\hat{\theta}$

$$\hat{\theta}^* = t(\boldsymbol{x}^*) \ .$$

The quantity $t(\boldsymbol{x}^*)$ is the result of applying the same function $t(\cdot)$ to $\boldsymbol{x}^*$ as was applied to $\boldsymbol{x}$.

Example 225 Continuing the example 224 we can draw a bootstrap sample as:

0.180	0.180	0.494	0.494	0.494	0.774	0.774	1.039	1.521	1.648
1.648	1.676	1.676	3.192	3.317	3.501	3.760	5.004	5.004	5.004

and

$$\hat{\theta}^* = 6.957 \ . \qquad \boxtimes$$

The boostrap estimate of $\mathrm{se}_{\hat{F}}(\hat{\theta})$, the standard error of a statistic $\hat{\theta}$, is a plug–in estimate that uses the empirical distribution function $\hat{F}$ in place of F. The bootstrap estimate of $\mathrm{se}_{\hat{F}}(\hat{\theta})$ is

$$\mathrm{se}_{\hat{F}}(\hat{\theta}^*)$$

This formula is called the **ideal bootstrap estimate of standard error** of $\hat{\theta}$.

Example 226 We use the bootstrap sample in example 225. Let μ be the first moment of a r.v. X. Then the plug–in principle leads to $\hat{\mu}$ as an estimate of μ. Further, simple calculation show

$$\mathrm{se}_{\hat{F}}(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$$

and the ideal bootstrap estimate of standard error is

$$\mathrm{se}_{\hat{F}}(\hat{\mu}^*) = \frac{\hat{\sigma}^*}{\sqrt{n}} = 0.366$$

and

$$\hat{\sigma}^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^* - \hat{\mu}^*)^2} = 1.636\ . \qquad \boxtimes$$

Unfortunately, for virtual any estimate $\hat{\theta}$ other than the mean, there is no neat formula for the standard error that enables us to compute the numerical value of the ideal estimate exactly.

The bootstrap algorithm, described next, is a computational way of obtaining a good approximation to the numerical value of $\mathrm{se}_{\hat{F}}(\hat{\mu}^*)$.

The Bootstrap algorithm for estimating standard errors

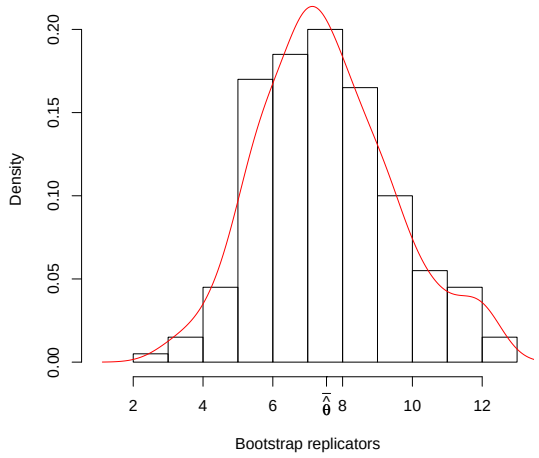
1. Select B independent bootstrap samples $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \cdots, \boldsymbol{x}_B^*$, each consisting of n data values drawn with replacement form $\boldsymbol{x}$ (For estimating a standard error, the number of B will ordinarily be in the range 25–200);

2. Evaluate the bootstrap replication corresponding to each bootstrap sample,
$$\hat{\theta}_b^* = t(\boldsymbol{x}_b^*) \qquad 1 \le b \le B\ ;$$

3. Estimate the standard error $\mathrm{se}_F(\hat{\theta})$ by the sample standard deviation of the B replications
$$\hat{\mathrm{se}}_B = \sqrt{\sum_{b=1}^{B}\frac{(\hat{\theta}_b^* - \bar{\hat{\theta}}^*)^2}{B-1}}$$
where $\bar{\hat{\theta}}^* = \sum_{b=1}^{B}\hat{\theta}_b^*/B$.

Example 227 In order to estimate $\mathrm{se}_F(\hat{\theta})$ when $\hat{\theta}$ is the second moment we set $B = 200$ and we obtain the bootstrap replications

7.286	6.791	4.351	7.607	7.276	7.430	7.354	6.412	8.942
5.300	10.073	5.664	9.400	10.460	4.775	6.282	9.690	5.878
6.581	3.625	5.756	5.499	7.538	7.055	5.956	9.267	4.891
11.421	5.296	6.911	8.165	8.662	6.506	10.633	10.101	10.969
11.363	5.775	6.816	7.762	8.836	6.440	8.619	7.205	8.066
11.997	9.712	6.745	5.853	6.880	8.623	5.102	10.470	2.906
8.380	5.773	9.364	4.316	9.135	5.443	8.044	6.598	7.493
5.893	6.967	8.684	7.150	3.309	6.169	7.575	5.653	5.545
8.278	6.912	8.131	8.205	9.393	11.779	8.810	7.871	5.973
6.114	9.824	7.327	12.108	11.659	8.095	7.419	7.184	8.144
5.414	9.252	9.491	6.077	5.428	8.190	6.415	5.941	5.109
8.349	6.412	10.516	4.252	7.116	9.086	7.511	8.804	10.148
4.762	6.713	7.764	5.607	6.689	8.409	6.756	8.728	8.173
6.938	6.998	5.988	11.995	9.232	6.374	10.809	6.888	7.359
11.083	7.225	8.670	8.035	9.535	6.243	6.153	6.551	10.328
7.060	6.675	3.472	7.231	9.299	6.588	6.586	5.117	7.548
6.313	7.448	11.943	8.426	5.725	10.262	7.415	7.976	8.236
7.976	9.557	6.972	12.082	9.218	5.975	4.609	9.352	6.647
7.018	7.700	7.285	9.678	8.255	6.965	7.684	5.915	4.125
9.228	7.351	5.270	7.034	8.274	7.859	5.699	6.921	8.124
6.784	5.585	7.728	8.870	12.334	5.426	11.932	7.644	4.371
9.073	8.947	7.985	7.005	5.039	8.739	8.030	5.468	7.324	5.139



Histogram and nonparametric kernel density estimator for the $B = 200$ bootstrap replications.

And then using step 3. we have

$$\bar{\hat{\theta}}^* = \sum_{b=1}^{B} \hat{\theta}_b^*/B = \frac{1507.725}{200} = 7.539$$

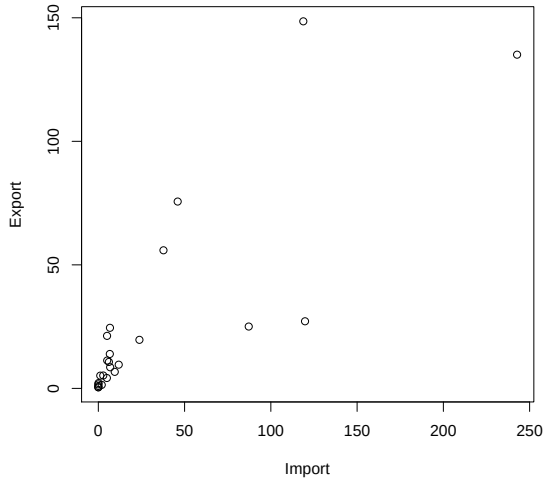
and

$$\hat{\text{se}}_B = \sqrt{\sum_{b=1}^{B} \frac{(\hat{\theta}_b^* - \bar{\hat{\theta}}^*)^2}{B-1}} = \sqrt{\frac{739.792}{199}} = 1.923 \qquad \boxtimes$$

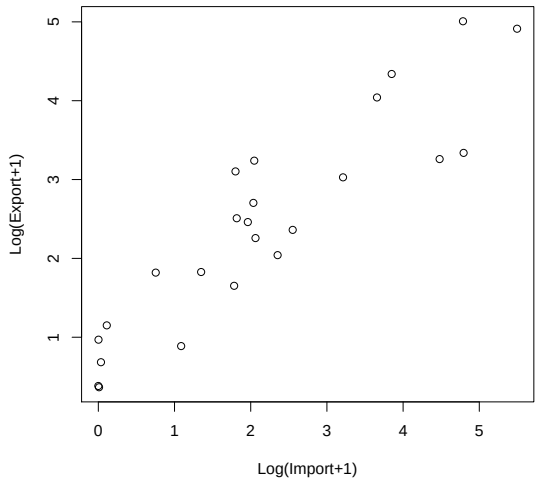
Example 228 (Import and Export of Italy)

Code	State	Import	Export
0001	Francia	118.82	148.57
0003	Paesi Bassi	87.18	25.04
0004	Germania	242.74	135.08
0006	Regno Unito	37.76	55.90
0007	Irlanda	1.12	5.17
0008	Danimarca	9.52	6.70
0009	Grecia	6.74	24.53
0010	Portogallo	5.15	11.29
0011	Spagna	46.00	75.63
0017	Belgio	119.82	27.17
0018	Lussemburgo	1.97	1.43
0030	Svezia	6.87	8.56
0032	Finlandia	2.85	5.22
0038	Austria	23.82	19.66
0046	Malta	0.12	2.16
0053	Estonia	0.00	0.47
0054	Lettonia	0.01	0.44
0055	Lituania	0.04	0.98
0060	Polonia	5.05	21.27
0061	Repubblica Ceca	6.65	13.94
0063	Slovacchia	4.94	4.22
0064	Ungheria	11.82	9.61
0091	Slovenia	6.11	10.71
0600	Cipro	0.00	1.64

Import and Export of Italy in EU25, December 2004, Plastic Material sector (NC8 classification number 39). wwww.coeweb.istat.it. Data set created Oct. 10, 2006.



Import and Export of Italy in EU25, December 2004, Plastic Material sector (NC8 classification number 39). wwww.coeweb.istat.it. Data set created Oct. 10, 2006.



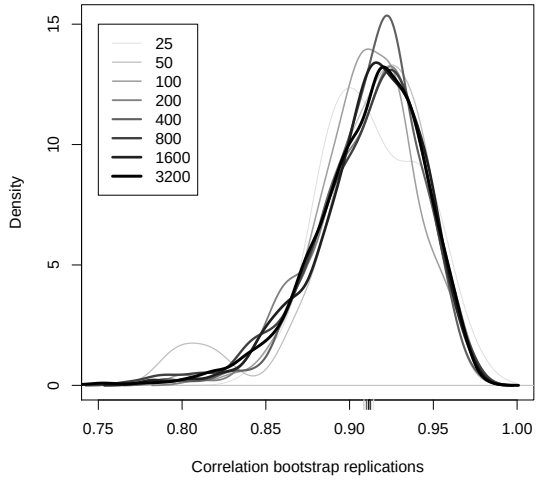
$\log(\text{Import} + 1)$ and $\log(\text{Export} + 1)$ of Italy in EU25, December 2004, Plastic Material sector (NC8 classification number 39). wwww.coeweb.istat.it. Data set created Oct. 10, 2006.

Given this dataset the correlation coefficient on the transformed data is

$$\text{cor}(\log(\text{Import}+1), \log(\text{Export}+1)) = 0.913$$

If we think about this dataset has a sample from all the import-export relation between Italy and EU25 state in December (pasts and futures) what is the standard deviation of this estimate?

B	25	50	100	200	400	800	1600	3200
$\hat{\text{se}}_B(\hat{\theta}^*)$	0.028	0.037	0.029	0.030	0.029	0.035	0.032	0.033



Nonparametric kernel density estimator of the correlation bootstrap replications for different number B of replications. ☒

The number of bootstrap replications B

How large should we take B, the number of bootstrap replications used to evaluate $\hat{\text{se}}_B$?

The ideal bootstrap estimate $\hat{\text{se}}_\infty$ takes $B = \infty$, in which case it equals the plug–in estimate

$$\hat{\text{se}}_\infty = \text{se}_{\hat{F}}(\hat{\theta}^*)$$

and hence $\hat{\text{se}}_\infty$ it has the smallest possible standard deviation among nearly unbiased estimates of $\text{se}_F(\hat{\theta})$, at least in an asymptotic ($n \to \infty$) sense.

$\hat{\text{se}}_B$ always has greater standard deviation than $\hat{\text{se}}_\infty$. The practical question is "how much greater"?

Two rules of thumb are as follow

- Even a small number of bootstrap replications, say $B = 25$, is usually informative. $B = 50$ is often enough to give a good estimate of $\mathrm{se}_F(\hat{\theta})$;
- Very seldom are more that $B = 200$ replications needed for estimating a standard error. (Much bigger values of B are required for bootstrap confidence intervals).

Mean while it pays to remember that bootstrap data, like real data, deserves a close look. In particular, it is almost never a waste of time to display the histogram of the bootstrap replications.

The parametric bootstrap

The parametric bootstrap uses an estimate $M(\hat{\theta})$ (from a parametric family) of the unknown distribution F instead of $\hat{F}$ in the bootstrap algorithm. In particular, the parametric bootstrap estimate of standard error is defined as

$$\mathrm{se}_{M(\hat{\theta})}(\hat{\theta}^*) \ .$$

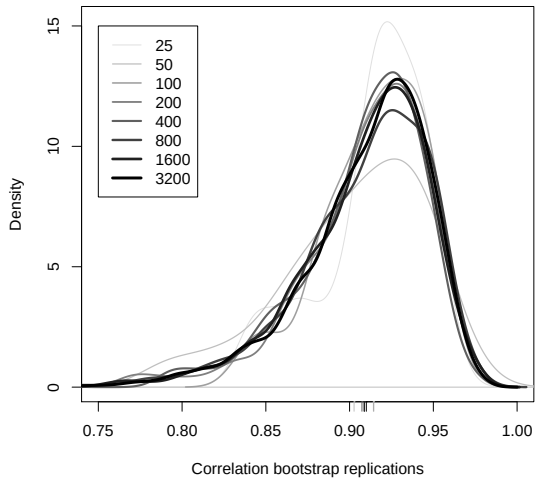
The rest of the algorithm is as before.

Example 229 Let assume that the joint distribution of the random variable $X = \log(\text{Import} + 1)$ and $Y = \log(\text{Export} + 1)$ is a bivariate normal distribution. Reasonable estimates of the mean and covariance of this population are given by

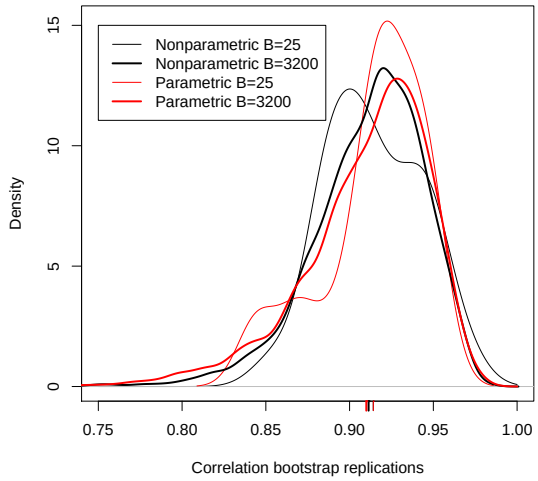
$$\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} = \begin{bmatrix} 2.168 \\ 2.431 \end{bmatrix}$$

$$\frac{1}{24} \begin{bmatrix} \sum (x_i - \bar{x})^2 & \sum (x_i - \bar{x})(y_i - \bar{y}) \\ \sum (x_i - \bar{x})(y_i - \bar{y}) & \sum (y_i - \bar{y})^2 \end{bmatrix} = \begin{bmatrix} 2.768 & 2.035 \\ 2.035 & 1.796 \end{bmatrix}$$

B	25	50	100	200	400	800	1600	3200
$\hat{\mathrm{se}}_B(\hat{\theta}^*)$	0.028	0.037	0.029	0.030	0.029	0.035	0.032	0.033
$\hat{\mathrm{se}}_{B,par}(\hat{\theta}^*)$	0.030	0.042	0.030	0.039	0.038	0.041	0.038	0.038



Nonparametric kernel density estimator of the correlation **parametric** bootstrap replications for different number B of replications.



Nonparametric kernel density estimator of the correlation **nonparametric** bootstrap replications and of the correlation **parametric** bootstrap replications for $B = 25$ and 3200. ☒

6.2 Confidence intervals

Bootstrap Confidence Intervals

There are several ways of construct confidence intervals using bootstrap:

- Confidence intervals based on bootstrap "tables" (bootstrap–t)
- Confidence intervals based on bootstrap percentiles (percentile interval)
- Bias–Corrected and accelerated method (BCa)
- Approximate Bootstrap Confidence intervals (ABC)

Here we are going to present only the first two. For the last ones see for instance Efron and Tibshirani [1993].

Bootstrap–t

In many situations the confidence interval is derived from a (quasi)–pivotal quantity

$$Z = \frac{\hat{\theta} - \theta}{\hat{\text{se}}} \ .$$

For some problems the distribution of Z is known exactly (for instance, if the data are iid from a normal distribution and θ is the location parameter then Z is a t distribution with the appropriate degrees of freedom). In general, the distribution would be unknown, or it may holds only approximately.

In the **bootstrap–t** approach the distribution of Z is estimated directly from the data: in essence it builds a table of quantiles for Z that is appropriate for the data set at hand.

We generate B bootstrap samples $\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_B^*$ and for each we compute

$$Z_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\hat{\text{se}}_b^*} \ .$$

where $\hat{\theta}_b^* = t(\boldsymbol{x}_b^*)$ is the value of $\hat{\theta}$ for the bootstrap sample $\boldsymbol{x}_b^*$ and $\hat{\text{se}}_b^*$ is the estimated standard error of $\hat{\theta}_b^*$ for the bootstrap sample $\boldsymbol{x}_b^*$.

The αth percentile of Z_b^* is estimated by the value z_α^* such that

$$\#\{Z_b^* \leq z_\alpha^*\}/B = \alpha$$

For example, if $B = 1000$, the estimate of the 5% point is the 50th largest value of the Z_b^*s. Finally the bootstrap–t confidence interval is

$$(\hat{\theta} - z_{(1-\alpha/2)}^* \hat{\text{se}} \ , \ \hat{\theta} - z_{\alpha/2}^* \hat{\text{se}}) \ .$$

- In the denominator of the Z_b^* there is $\hat{\text{se}}_b^*$ which is the bootstrap estimated standar error (at least when there is no closed formula for it) for the given bootstrap sample $\boldsymbol{x}_b^*$. This means, for every (main) bootstrap sample $\boldsymbol{x}_b^*$ we have to run a bootstrap algorithm in order to estimate its standard error. That is, a nested two layer bootstrap algorithm is used.

- If $B \times \alpha$ is not an integer, the αth percentile of Z_b^* must be evaluated with (linear) approximation (as it is done for the evaluation of the quantile in a discrete distribution).

- The bootstrap–t interval may performs erratically in small-sample, in the nonparametric settings.

7 Stability of Inference

7.1 Introduction

7.2 Measures of stability

7.3 Multicollinearity

7.4 Ridge regression

7.5 Robust Estimation

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. B.N. Petrov and F. Csáki, editors, *2nd International Symposium of Information Theory*, pages 267–281, Akadémiai Kiadó, Budapest, 1973.

A. Azzalini. *Statistical Inference Based on the Likelihood*. Chapman & Hall, 1996.

R.A. Beezer. *A First Course in Linear Algebra*. on line, version 0.80, august 22, 2006 edition, 2006. URL `linear.ups.edu`.

L. Boltzmann. über die beziehung zwischen dem hauptsatze derzwe: Ten mechanischen wärmetheorie und der wahrscheinlichkeitsrechnung respective den sätzen über das wärmegleichgewicht. *Wiener Berichte*, 76:373–435, 1877.

D. Bystrak. Evaluation of some random effects methodology applicable to bird ringing data. In C.J. Ralph and J.M. Scott, editors, *Estimating numbers of terrestrial birds. Studies in Avian Biology*, volume 6, pages 522–532. 1981.

J. Cavanaugh and R. Shumway. A bootstrap variant of aic for state-space model selection. *Statistica Sinica*, 7:473–496, 1997.

J.E. Cavanaugh and A.A. Neath. Generalizing the derivation of the schwarz information criterion. *Communications in Statistics -Theory and Methods*, 28:49–66, 1999.

S. Chatterjee and A.S. Hadi. *Sensitivity analysis in linear regression*. Wiley, New York, 1988.

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

A. Davison. *Statistical Models*. Cambridge University press, 2003. ISBN 0–521–77339–3.

B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979.

B. Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382): 316–331, 1983.

B. Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81:461–470, 1986a.

B. Efron. Discussion of the paper by c.f.j. wu. *Annals of Statistics*, 14: 1301–1304, 1986b.

B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1993.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, 2001. ISBN 0–387–95187–3.

C.H. Flather. *Pattern of avian species–accumulation rates among easter forested landscapes*. PhD thesis, Colorado State University, Fort Collins, CO, 1992.

C.H. Flather. Fitting species–accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography*, 23:155–168, 1996.

F.A. Graybill. *An Introduction to Linear Statistical Models*, volume I. McGraw–Hill, 1961.

D.M.A. Haughton. On the choice of a model to fit data form an exponential family. *Annals of Statistics*, 6:342–355, 1988.

C.M. Hurvich and C-L. Tsai. Regresson and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

M. Ishiguro and Y. Sakamoto. Information criterion and bootstrap method. In Japan Statist. Soc., editor, *Proceedings of the Annual Meeting of Japan Statistical Society*, pages 156–158, 1991.

S. Kullback and R.A. Leibler. On informaton and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

T. Leonard. Comment on "a simple predictive density function" by m. lejeune and g. d. faulkenberry. *Journal of the American Statistical Association*, 77 (370):657–658, 1982.

A.J. Miller. *Subset Selection in Regression*. Chapman & Hall, 2006. ISBN 1–584–88171–2.

D.C. Montgomery and E.A. Peck. *Introduction to linear regression analysis*. John Wiley, New York, 1982.

Y. Pawitan. *In All Likelihood*. Oxford Science Publications, 2001.

R.D. Martin R.A. Maronna and V.J. Yohai. *Robust Statistics*. Wiley, 2006. ISBN 0–470–01092–4.

C.R. Rao and H. Toutenburg. *Linear Models*. Springer, 1995. ISBN 0–387–94562–8.

H. Scheffé. *The Analysis of Variance*. Wiley-Interscience, 1999. ISBN 0–471–34505–9.

J.R. Schott. *Matrix analysis for statistics*. Wiley, 1997. ISBN 0–471–15409–1.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6: 461–474, 1978.

C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

R. Shibata. Statistical aspects of model selection. In J.C. Willems, editor, *From data to model*, pages 215–240. Springer–Verlag, 1989.

R. Shibata. Bootstrap estimate of kullback–leibler information for model selection. *Statistica Sinica*, 7:375–394, 1997.

M. Stone. Comments on model selection criteria of akaike and schwarz. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2): 276–278, 1979.

R Core Team. *An Introduction to R.* on line, version 2.3.1 (2006-06-01) edition, 2006. ISBN 3–900051–12–7. URL `cran.r-project.org`.

W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S.* Springer, fourth edition, 2002. ISBN 0–387–95457–0.

Index

8 GNU Free Documentation License

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The **"Document"**, below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as **"you"**. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A **"Modified Version"** of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A **"Secondary Section"** is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The **"Invariant Sections"** are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The **"Cover Texts"** are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A **"Transparent"** copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called **"Opaque"**.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The **"Title Page"** means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License

requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section **"Entitled XYZ"** means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as **"Acknowledgements"**, **"Dedications"**, **"Endorsements"**, or **"History"**.) To **"Preserve the Title"** of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title

equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled ”Endorsements” or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled ”Endorsements”, provided it contains nothing but endorsements of your Modified Version by various parties–for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section

titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in

addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.