ECE 645 – Matrix Equations and Least Squares Estimation

J. V. Krogmeier

March 24, 2014

Contents

1 Matrix Equations 4

2 Exact Solution of Linear Equations 6

3 Inner Product and Norm 8

4 The Projection Theorem and Related 9

5 Eigenvalues and Eigenvectors 12
6 Projection Matrices
7 Construction of Orthogonal Projectors
8 An Adjoint Theorem
9 Left and Right Inverses
10 Some Theorems on the Solution of $Ax = b$
   10.1 Over-determined Full Rank Case
   10.2 Under-determined Full Rank Case
   10.3 The General Case
11 The Singular Value Decomposition
12 Deterministic Least Squares Problem
13 Statistical Least Squares and the Additive Linear Model
14 Identifiability in the Additive Linear Model
15 The Gauss-Markov Theorem

15.1 Reprise of Least Squares ......................................................... 34
15.2 Usually Least Squares is Not the Best We Can Do ................. 35
1 Matrix Equations

1. Consider a matrix equation of the form

\[
\begin{bmatrix}
a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\
a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m,1} & a_{m,2} & \cdots & a_{m,n}
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{bmatrix} = 
\begin{bmatrix}
b_1 \\
b_2 \\
\vdots \\
b_m
\end{bmatrix}.
\]

The usual shorthand for the linear system of equations above is

\[Ax = b\] (1)

where \(A\) is an \(m \times n\) matrix, \(x\) is an \(n \times 1\) vector, and \(b\) is an \(m \times 1\) vector.

**Goal:** To study the exact and approximate solutions to linear systems as in Eq. (1).

In the process we will make extensive use of the ideas of orthogonality, projection, and eigenvalue decomposition. Introductory treatments of these topics consume entire volumes so no attempt will be made for completeness. See any of a number of good texts on linear algebra [1, 2, 3, 4]

2. Assume familiarity with the following concepts (see the first four chapters of [1]):

(a) The usual rules for matrix addition and multiplication.
(b) The notions of a basis for a vector space, change of basis, and the matrix representation of a linear transformation with respect to a particular basis. Similarity transformations and diagonalization.

(c) Transpose $A^T$ and Hermitian transpose $A^H$ of a matrix $A$.

(d) Determinant of a matrix $|A|$, the rank of a matrix, row rank, column rank, the inverse of a square matrix.

(e) The standard orthonormal basis of the vector spaces $\mathbb{R}^n$ or $\mathbb{C}^n$ is the collection of $n$ vectors

$$\{e_j : 1 \leq j \leq n\}$$

where $e_j$ denotes the $n$-vector whose only nonzero entry is a one in the $j$-th position, i.e.,

$$e_j = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow j^{th}.$$
2 Exact Solution of Linear Equations

1. Let $A$ be an $m \times n$ matrix with elements from $\mathbb{C}$. The null space of $A$ is the following subspace of $\mathbb{C}^n$

$$\eta(A) = \{x \in \mathbb{C}^n : Ax = 0\}.$$ 

The range space or column space of $A$ is the following subspace of $\mathbb{C}^m$

$$\mathcal{R}(A) = \{y = Ax : x \in \mathbb{C}^n\},$$

i.e., that subspace spanned by the column vectors of $A$. The (column) rank of the matrix $A$ is defined to be the dimension of its range space.

2. It is an important and interesting fact\(^1\) that the rank of $A$ is also equal to the dimension of the column space of $A^T$ (or, equivalently, the row space of $A$).

3. Some simple theorems:

**Thm 1.** Let $A$ be an $m \times n$ matrix. Then

$$\text{rank}(A) + \dim\{\eta(A)\} = n.$$

\(^1\)Another notion for the rank of a rectangular matrix is that of the order of its largest nonzero minor [1]. All of these, row rank, column rank, and determinental rank, turn out to be equal.
Thm 2. Let $A$ be an $m \times n$ matrix and $b$ be an $m$-vector, both considered as constant and known. Then considering the $n$-vector solutions $x$ to the matrix equation $Ax = b$ one can say that

(a) there exists a solution $x$ if and only if $b \in \mathcal{R}(A)$.

(b) if $x^o$ is a particular solution of the matrix equation, then the complete set of solutions to Eq. (1) is given by

$$\{x^o + x^\eta : x^\eta \in \eta(A)\}.$$ 

(c) a solution is unique only if $\eta(A) = \{0\}$.

(d) the matrix equation has a unique solution for any $b$ if and only if $n = m$ and $\text{rank}(A) = n$. In this case, the unique solution is given by

$$x = A^{-1}b$$

where $A^{-1}$ is the (unique) inverse matrix for $A$. It solves the matrix equation

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

Thm 3. Let $A$ and $B$ be any two matrices such that $AB$ is defined. Then

$$\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}.$$
**Thm 4.** Let \( A \) be an \( m \times n \) matrix with nonzero rank equal to \( r \). Then there exist nonsingular matrices \( U (m \times m) \) and \( V (n \times n) \) such that

\[
A = U \begin{bmatrix}
I_{r \times r} & 0_{r \times (n-r)} \\
0_{(m-r) \times r} & 0_{(m-r) \times (n-r)}
\end{bmatrix} V.
\]

**3 Inner Product and Norm**

1. In the vector space \( \mathbb{C}^n \) the standard **inner product** of two vectors \( x \) and \( y \) is defined as

\[
\langle x, y \rangle = y^H x = \sum_{k=1}^{n} x_k y_k^*.
\]

   This inner product defines a norm

\[
\|x\| = \sqrt{\langle x, y \rangle}.
\]

2. Two vectors \( x \) and \( y \) are said to be **orthogonal** with respect to the standard inner product if \( \langle x, y \rangle = 0 \). This may be used to define orthogonal subsets and subspaces [1, page 111]. If \( \mathcal{V} \) is a subspace of \( \mathbb{C}^n \) we define

\[
\mathcal{V}^\perp = \{ w \in \mathbb{C}^n : \langle w, v \rangle = 0 \text{ for all } v \in \mathcal{V} \}
\]

   and call it the **orthogonal complement** of \( \mathcal{V} \).
3. There are many variations on this theme, see [1, pages 104-113]. In fact, it can be shown that the most general inner product on $\mathbb{C}^n$ can be written as

$$\langle x, y \rangle_Q = y^H Q x$$

where $Q$ is an $n \times n$ positive definite matrix$^2$.

**Example.**
Blah blah.

4. **The Projection Theorem and Related**

1. The following result is essential to all that follows. It is commonly known as the **Projection Theorem**. We only state it in the generality required here. For more information see [5].

**Thm 5.** [Projection Theorem] Let$^3$ $\mathcal{H}$ be a finite dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$ with an inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. Let $\mathcal{M}$ be a subspace of $\mathcal{H}$.

$^2$A square matrix $Q$ is said to be **positive definite** if $x^H Q x > 0$

for all nonzero vectors $x$.

$^3$ i.e., it is either $\mathbb{R}^n$ or $\mathbb{C}^n$ and the inner product is of the form $\langle x, y \rangle = y^H Q x.$
Then for any $x \in \mathcal{H}$ there exists a unique vector $y_* \in \mathcal{M}$ such that

$$\|x - y_*\| \leq \|x - y\|$$

for all $y \in \mathcal{M}$. Moreover, the unique vector $y_*$ is characterized by the orthogonality condition

$$\langle x - y_*, y \rangle = 0$$

for all $y \in \mathcal{M}$.

Proof.

**Thm 6.** [Dual Projection Theorem] Let $\mathcal{H}$ be a finite dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$ with an inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. Let $\mathcal{M}$ be a subspace, $x \in \mathcal{H}$, and define the linear variety

$$\mathcal{V} = x + \mathcal{M} = \{x + y : y \in \mathcal{M}\}.$$ 

Then there is a unique vector $v_* \in \mathcal{V}$ of smallest norm, i.e., such that

$$\|v_*\| \leq \|v\|$$

for all $v \in \mathcal{V}$. Moreover, $v_* \in \mathcal{M}^\perp$.

Proof.
Thm 7. [Corollary to Dual Projection Theorem] Let \( \mathcal{H} \) be finite dimensional over \( \mathbb{R} \) or \( \mathbb{C} \) and suppose that \( \{y_1, y_2, \ldots, y_m\} \) is a linearly independent set of vectors from \( \mathcal{H} \). Then among all vectors \( x \in \mathcal{H} \) satisfying the constraints

\[
\langle x, y_1 \rangle = c_1 \\
\langle x, y_2 \rangle = c_2 \\
\vdots \\
\langle x, y_m \rangle = c_m
\]

there is a unique \( x_0 \) having smallest norm. It is given by

\[
x_0 = \sum_{k=1}^{m} \beta_k y_k
\]

where the \( \beta_k \) solve \( G\beta = c \) in terms of the so-called Gram matrix

\[
G = G(y_1, y_2, \ldots, y_m) = \begin{bmatrix}
\langle y_1, y_1 \rangle & \langle y_2, y_1 \rangle & \cdots & \langle y_m, y_1 \rangle \\
\langle y_1, y_2 \rangle & \langle y_2, y_2 \rangle & \cdots & \langle y_m, y_2 \rangle \\
\vdots & \vdots & \cdots & \vdots \\
\langle y_1, y_m \rangle & \langle y_2, y_m \rangle & \cdots & \langle y_m, y_m \rangle
\end{bmatrix}
\]

of the \( y_k \).

Proof.
5 Eigenvalues and Eigenvectors

1. A a square matrix, say \( n \times n \). A nonzero vector \( v \) is said to be an eigenvector of \( A \) corresponding to the eigenvalue \( \lambda \in \mathbb{C} \) if

\[
Av = \lambda v.
\]

2. The eigenvalues are the solutions to the polynomial equation (the characteristic polynomial)

\[
|\lambda I - A| = 0.
\]

3. The following properties are often useful. They apply to a square matrix \( A \), which is Hermitian symmetric, i.e., \( A^H = A \).

(a) \( A \) has real eigenvalues.

(b) There exists an orthonormal basis for \( \mathbb{C}^n \) consisting of eigenvectors of \( A \), in particular, \( A \) is diagonalizable.

(c) In any case the eigenvectors of \( A \) corresponding to distinct eigenvalues are orthogonal.

(d) If, in addition, \( A \) is positive definite then its eigenvalues are positive. If \( A \) is nonnegative definite then its eigenvalues are nonnegative.
See [1, pages 147-160] for further details and additional properties.

6 Projection Matrices

1. **Projection matrices** are square matrices satisfying the idempotency condition: \( P^2 = P \). For such matrices we have the following facts [1, pages 194-199].

**Thm 8.** If \( P \) is idempotent, then:

(a) \( I - P \) is idempotent.
(b) \( \mathcal{R}(I - P) = \eta(P) \).
(c) \( \eta(I - P) = \mathcal{R}(P) \).

Note that the idempotent matrix \( Q = I - P \) is called the **complementary projector** to \( P \).

**Thm 9.** If \( P \) is an \( n \times n \) projection matrix, then\(^4\) \( \eta(P) + \mathcal{R}(P) = \mathcal{C}^n \), where \( \eta(P) \cap \mathcal{R}(P) = \{0\} \).

2. Blah.

\(^4\)Blah.
3.

**Thm 10.** Let \( P \) be an \( n \times n \) projection matrix. Then the eigenvalues of \( P \) are all either 0 or 1 and, furthermore, \( P \) is diagonalizable. That is, there exists a nonsingular \( n \times n \) matrix \( B \) such that

\[
P = B \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} B^{-1}.
\]

*Proof.*

7 Construction of Orthogonal Projectors

1. Note that in the diagonal representation of the projection \( P \) of Thm 10, the columns of the nonsingular matrix \( B \) are eigenvectors of \( P \).

2. Furthermore, if \( P \) is Hermitian symmetric, these eigenvectors may be chosen to be orthonormal, i.e., we may assume that \( B^{-1} = B^H \).

3. Then with

\[
B = \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix}
\]
written in terms of its orthonormal column vectors, we may write
\[
P = B \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} B^{-1} = \sum_{j=1}^{r} b_j b_j^H.
\]

4. The previous shows that the action of an orthogonal projector on an arbitrary vector \( x \in \mathbb{C}^n \) can be written
\[
P x = \sum_{j=1}^{r} b_j b_j^H x = \sum_{j=1}^{r} \langle x, b_j \rangle b_j.
\]
Thus, each term in the sum on the right is the orthogonal projection of the vector \( x \) onto the subspace spanned by one of the eigenvectors of \( P \).

5. Now consider a more direct approach to the construction of orthogonal projection matrices. Suppose that one is given a subspace \( \mathcal{V} \subset \mathbb{C}^n \) with a basis comprising the columns of an \( n \times k \) matrix
\[
V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}.
\]

6. Clearly a typical member of \( \mathcal{V} \) is of the form \( V x \) for some \( x \in \mathbb{C}^k \). Given a vector \( w \in \mathbb{C}^n \) we would like to find a \( V x \) that can be called the projection of \( w \) onto \( \mathcal{V} \).

7. The condition characterizing such a \( V x \) is that \( w - V x \) be orthogonal to \( \mathcal{V} \). Thus, for the standard inner product
\[
v_j^H (w - V x) = 0 \quad \text{for} \quad 1 \leq j \leq k.
\]
This is equivalent to the matrix equation

\[ V^H(w - Vx) = 0. \]

8. Solving it: \( x = (V^H V)^{-1} V^H w. \)

9. Hence the desired orthogonal projection matrix is

\[ P = V(V^H V)^{-1} V^H. \]

8. **An Adjoint Theorem**

1.

**Thm 11.** Let \( A \) be an \( m \times n \) matrix and \( A^H \) the \( n \times m \) complex conjugate transpose. Then

(a) \( \text{rank}(A) = \text{rank}(A^H) \).

(b) Sometimes the following is called the Fundamental Theorem of Linear Algebra

\[
\begin{align*}
\mathcal{C}^n &= \eta(A) \oplus \mathcal{R}(A^H) \\
\mathcal{C}^m &= \eta(A^H) \oplus \mathcal{R}(A).
\end{align*}
\]
(c) $\mathcal{R}(A^H A) = \mathcal{R}(A^H)$ and $\eta(A^H A) = \eta(A)$.
(d) $\mathcal{R}(AA^H) = \mathcal{R}(A)$ and $\eta(AA^H) = \eta(A^H)$.
(e) $A$ is invertible if and only if $A^H$ is invertible and
\[
(A^H)^{-1} = (A^{-1})^H.
\]

Proof.

9 Left and Right Inverses

1. A matrix has an inverse if and only if it is square and of full rank. This also turns out to be equivalent to the existence and uniqueness of the solutions of Eq. (1) for an arbitrary right side vector $b$.

2. It often happens that a matrix equation does not have an invertible coefficient matrix in which case a more general notion of matrix inverse helps to characterize the solutions to Eq. (1).

3. Let $A$ be an $m \times n$ matrix. Then $A$ is said to be

   (a) **left invertible** if there exists an $n \times m$ matrix $A^L$ such that $A^L A = I_n$.

   (b) **right invertible** if there exists an $n \times m$ matrix $A^R$ such that $AA^R = I_m$.
4. The following facts give necessary and sufficient conditions for the left or right invertibility of $A$.

**Thm 12.** The following statements are equivalent for any $m \times n$ matrix $A$:

(a) $A$ is left invertible;
(b) $m \geq n$ and rank($A$) = $n$;
(c) The columns of $A$ are linearly independent in $C^m$;
(d) $\eta(A) = \{0\}$.

**Thm 13.** The following statements are equivalent for any $m \times n$ matrix $A$:

(a) $A$ is right invertible;
(b) $m \leq n$ and rank($A$) = $m$;
(c) The rows of $A$ are linearly independent in $C^n$;
(d) $R(A) = C^m$.

5. In summary, an $m \times n$ matrix $A$ is one-sided invertible if and only if it has full rank. If this is the case and $m \geq n$ then $A$ is left invertible, if $m \leq n$ then $A$ is right invertible.
6. Left or right inverses are usually not unique. However, if a square matrix $A$ has a left inverse, then it also has a right inverse and the two inverses are equal and equal to $A^{-1}$. It is possible to write down a formula of sorts for all possible left (or right) inverses. See [1, Theorems 3, 4 on page 426].

**Example.**

7. Left and right inverses can be related to projection matrices. Suppose, for example, that $A$ is $m \times n$ and left invertible. First, we note that $AA^L$ is a left identity for $A$ because

$$(AA^L)A = A.$$ 

8. Second, we claim that $AA^L$ is idempotent and hence, it is a projection matrix onto some subspace along some direction. To see this just compute

$$(AA^L)(AA^L) = AI_nA^L = AA^L.$$ 

**Thm 14.** Suppose that $A$ is $m \times n$ and left invertible. Then for any left inverse $A^L$

(a) $\mathcal{R}(A) = \mathcal{R}(AA^L)$ and $\eta(AA^L) = \eta(A^L)$.
(b) $AA^L$ is a projector onto $\mathcal{R}(A)$.
(c) $\dim\{\eta(A^L)\} = m - \dim\{\mathcal{R}(A)\}$. 

Matrices and Least Squares 19
Proof.

9. Similar things can be said about right invertibility. Namely, for any right inverse $A^R$ of a matrix $A$, $A^RA$ is a right identity for $A$

$$A(A^RA) = A$$

and $A^RA$ is idempotent and hence is a projection matrix. Furthermore, one has the following result.

**Thm 15.** Suppose that $A$ is $m \times n$ and right invertible. Then for any left inverse $A_R$

(a) $\mathcal{R}(A^R) = \mathcal{R}(A^RA)$ and $\eta(A^RA) = \eta(A)$.

(b) $I_n - A^RA$ is a projector onto $\eta(A)$.

(c) $\dim \{ \mathcal{R}(A^R) \} = n - \dim \{ \eta(A) \}$.

10. The next two results have a direct bearing on the solutions of Eq. (1). We do not prove them here because they are a special case of a more general result which will follow.

**Thm 16.** If $A$ is left invertible, then the equation $Ax = b$ if solvable if and only if

$$(I_m - AA^L)b = 0.$$ (2)
When Eq. (2) hold the solution of $Ax = b$ is unique and given by

$$x = A^Lb.$$  

**Thm 17.** If $A$ is right invertible, then the equation $Ax = b$ if solvable for any $b \in \mathbb{C}^m$ and if $b \neq 0$, every solution $x$ is of the form

$$x = A^Rb$$

for some right inverse of $A$.

**Thm 18.** Let $A$ be $m \times n$ and suppose that $X$ is an $n \times m$ solution of

$$AXA = A. \quad (3)$$

Then the equation $Ax = b$ has a solution if and only if $AXb = b$ in which case any solution has the form

$$x = Xb + (I_n - XA)y \quad (4)$$

for some $y \in \mathbb{C}^n$.

**Proof.**

11. Theorem 18 includes the results of Theorems 16 and 17 as a special case.
(a) To see this suppose that $A$ is left invertible and say $A^L A = I_n$. Then it is easy to show that Eq. (3) has a solution $X$ which may be taken to be any left inverse $A^L$ of $A$. In this case the solvability condition (2) of Theorem 16 is equivalent to the solvability condition $AXb = b$ of Theorem 18. Moreover, by left invertibility the characterization (4) reduces to a unique solution $x = A^L b$.

(b) Now suppose that $A$ is right invertible and say $AA^R = I_m$. Now Eq. (3) has a solution $X$ which may be taken to be any right inverse $A^R$ of $A$. In this case the solvability condition of Theorem 18 is always satisfied. Looking at (4) we see that it reduces to

$$x = A^R b + (I_n - A^R A)y.$$ 

This can be simplified, however. Note that $A(I_n - A^R A)y = 0$ for all $y \in \mathbb{C}^n$. Therefore, the solutions $x$ might as well be parameterized by

$$x = A^R b$$

as suggested in Theorem 17.
10 Some Theorems on the Solution of $Ax = b$

10.1 Over-determined Full Rank Case

**Thm 19.** Let $A$ be an $m \times n$ matrix with $m \geq n = \text{rank}(A)$ and let $b \in \mathbb{C}^m$. Then there exists a unique $\hat{x}_{LS} \in \mathbb{C}^n$ such that

$$\|A\hat{x}_{LS} - b\| \leq \|Ax - b\|$$

for all $x \in \mathbb{C}^n$. The minimizer $\hat{x}_{LS}$ is given by

$$\hat{x}_{LS} = A^{LS}b$$

where

$$A^{LS} = (A^HA)^{-1}A^H.$$ 

Moreover, we note that

(a) $A\hat{x}_{LS} = (AA^{LS})b$ is the orthogonal projection of $b$ onto $\mathcal{R}(A)$.
(b) $A^{LS}$ is the unique left inverse of $A$ such that $\eta(A^{LS}) = \eta(A^H)$.
(c) $A^{LS}$ is the unique left inverse of $A$ such that $AA^{LS}$ is an orthogonal projection.

**Proof.**

1.
10.2 Under-determined Full Rank Case

**Thm 20.** Let $A$ be an $m \times n$ matrix with $n \geq m = \text{rank}(A)$ and let $b \in \mathbb{C}^m$. Then amongst all $x \in \mathbb{C}^n$ satisfying $Ax = b$ there exists a unique $\hat{x}_{MN} \in \mathbb{C}^n$ with smallest norm given by

$$\hat{x}_{MN} = A^H(AA^H)^{-1}b = A^{MN}b.$$ 

Moreover, we note that

(a) For any $x$ such that $Ax = b$, $\hat{x}_{MN} = A^{MN}Ax$ is the orthogonal projection of $x$ onto $\mathcal{R}(A^H)$.
(b) $A^{MN}$ is the unique right inverse of $A$ such that $\mathcal{R}(A^{MN}A) = \mathcal{R}(A^H)$.
(c) $A^{MN}$ is the unique right inverse of $A$ such that $A^{MN}A$ is an orthogonal projection matrix.

**Proof.**

1.

2.
10.3 The General Case

**Thm 21.** Let $A$ be an $m \times n$ matrix with $n, m \geq \text{rank}(A)$ and let $b \in \mathbb{C}^m$. Among the vectors $x$ that minimize $\|Ax - b\|$ the one with the least norm is unique and is the only vector in $\mathcal{R}(A^H)$ that minimizes $\|Ax - b\|$. In addition, it is the only vector that satisfies

$$A^H A A^H z = b$$
$$\hat{x}_{MNLS} = A^H z.$$

**Proof.**

1. 
2. 

11 The Singular Value Decomposition

**Thm 22.** Every $m \times n$ matrix $A$ of rank $r$ can be transformed into the form

$$A = R \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} Q^H$$
where $R^H R = RR^H = I_m$, $Q^H Q = QQ^H = I_n$, and $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_r\}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$.

Proof.

1.

2.
12 Deterministic Least Squares Problem

1. A deterministic procedure (not involving densities).

2. Can be analyzed and justified by probabilistic arguments which we will do later.

3. Provides a concrete introduction to basic estimation ideas \textit{and} is a practical technique.

4. The problem. Let $H$ be an $m \times n$ matrix and $y$ an $m$ vector. Solve the system of equations

$$ H\theta = y $$

for the $n$ vector $\theta$.

5. Consider the case $m > n$, i.e., the \textbf{overdetermined} case, where there are more equations than unknowns. There may not be a solution $\theta$ for given $H$ and $y$. In such a case we try to minimize the norm of the error\footnote{Also called the residual.}: $r = r(\theta) = y - H\theta$.

$$ \|r\|^2 = (y - H\theta)^T(y - H\theta) = y^Ty - y^TH\theta - \theta^TH^Ty + \theta^TH^TH\theta. $$

6. Take the gradient to find a stationary point of the error $\|r(\theta)\|^2$ as a function of $\theta$:

$$ \frac{\partial}{\partial \theta} \left\{ \|r(\theta)\|^2 \right\} = \begin{bmatrix} \frac{\partial \|r(\theta)\|^2}{\partial \theta_1} & \frac{\partial \|r(\theta)\|^2}{\partial \theta_2} & \ldots & \frac{\partial \|r(\theta)\|^2}{\partial \theta_n} \end{bmatrix}^T $$
\[ = -2H^T y + 2H^T H\theta \]
\[ = -2H^T (y - H\theta) \]
\[ = 0 \]

7. A value of \( \theta \) minimizing the norm of the residual is called a **least squares solution**. Also note that the Hession

\[
\frac{\partial}{\partial \theta} \left( \frac{\partial \|r(\theta)\|^2}{\partial \theta} \right)^T = 2H^T H \geq 0
\]

is non-negative definite, which means that the stationary point is a minimum for the norm of the residual\(^6\).

8. From the gradient equation we can say that a least squares solution must solve the square \( (n \times n) \) system

\[ H^T H \theta = H^T y \]

known as the **Normal equations**.

9. If \( H \) has rank \( n \) then there is a unique least squares solution given by

\[ \hat{\theta}_{LS} = (H^T H)^{-1} H^T y. \]

---

\(^6\)If \( H \) has full column rank then the Hession is actually positive definite.
13 Statistical Least Squares and the Additive Linear Model
Least Squares and Additive Linear Model

\[ \begin{align*}
\text{parameters} & \quad \begin{array}{ccc}
\Theta & \rightarrow & Y \\
\text{n vector} & \quad \text{(unknown d, deterministic)} & \quad \text{m vector}
\end{array}
\end{align*} \]

\[ Y = H\Theta + N \quad \text{also assume } m \gg n = \text{rank}(H). \]

Not knowing anything better to do it would seem reasonable (if \( n \) is somehow small) to estimate the true parameter \( \Theta \) given an observation \( Y \) by the least squares solution to the linear system \( H\Theta = Y \).

As we have seen, \( \hat{\Theta}_{ls} = \hat{\Theta}_{ls}(y) = (H^TH)^{-1}H^T \) is such that the residual

\[ r = r(\hat{\Theta}_{ls}, y) = y - H\hat{\Theta}_{ls} \]

has the smallest norm over \( \Theta \in \mathbb{R}^n \) (i.e., \( \| r(\hat{\Theta}_{ls}, y) \| \leq \| r(\Theta) \| \) ). When we put the random variable back in, then \( \hat{\Theta}_{ls} \) becomes a random variable itself:

\[ \hat{\Theta}_{ls}(\gamma) = (H^TH)^{-1}H^TY \]

One of our main objectives will be to consider the performance of this least squares estimator. In other words, what properties of the noise make this a good method or not.

Some basic observations/properties:

1. \( \hat{\Theta}_{ls}(Y) \) is a linear function of observations (linear estimator).
2. Estimation error \( \tilde{\Theta}_{ls} = \Theta - \hat{\Theta}_{ls} \) is a linear function of \( N \).
3. Note that \( \text{cov}(\tilde{\Theta}_{ls}) \) does not depend upon \( Y \).

\[ \text{Cov}(\tilde{\Theta}_{ls}) = E\{\tilde{\Theta}_{ls}\tilde{\Theta}_{ls}^T\} = E\{\tilde{\Theta}_{ls}\tilde{\Theta}_{ls}^T\} = \text{cov}(N) \]

This is a very useful property. Generally it is true for unbiased estimators.

In the special case where \( \text{cov}(N) = \sigma^2 I \) and \( E[\gamma] = 0 \), components of noise are uncorrelated and with identical variances.

\[ \text{Cov}(\tilde{\Theta}_{ls}) = \sigma^2 (H^TH)^{-1} \]
Covariance of residual given by

\[ \text{Cov}(rr^T) = \text{E}\{rr^T\} = [I_m - H(H^T H)^{-1} H^T] \text{Cov}(u) [I_m - H(H^T H)^{-1} H^T] \]

Since \( I_m - H(H^T H)^{-1} H^T \) is the orthogonal projection onto \( \eta(H^T) \), a space of dimension \( m-n \). It thus has rank \( m-n \) and so the cov. matrix \( \text{Cov}(r) \) of dimension \( m \times m \) has only rank \( m-n \). It is singular.

**Weighted Least Squares**

In the deterministic or random formulations we had used as error measure

\[ \|r(e)\|^2 = r^T r = (y - H\theta)(y - H\theta) \]

Given any symmetric positive definite matrix \( W \) (weighting matrix) we can generalize the problem to:

\[ \|r(e)\|^2 = r^T W r = (y - H\theta)^T W (y - H\theta) \]

*This amounts to the definition of a different inner prod.
and hence a different norm on \( \mathbb{R}^m \).*

*W can be used to emphasize or deemphasize the influence
of specific measurements in \( Y = [Y_1, Y_2, \ldots, Y_m]^T \).*

*For example, if \( Y_1 \) more reliable than \( Y_2 \) more reliable
than \( Y_3 \) \ldots etc. Could e.g. pick

\[ W = \text{diag}\{w_1, w_2, \ldots, w_m\} \quad w_1 > w_2 > w_3 \geq \cdots \geq w_m > 0 \]

**Simplifying Fact**

\[ W > 0 \implies W = V V^T \quad \text{(square root factorization)} \]

where \( V \) is nonsingular.

This can be used to reduce to a standard least squares
problem.

\[ \tilde{y} = V^T Y = V^T \theta + V^T N \]

\[ \tilde{y} = \theta + \nu \]

Then may easily see.

\[ \|r(e)\|^2_w = r^T W r = (\tilde{y} - \theta)^T W (\tilde{y} - \theta) \]
The least squares solution under full column rank assumption would be:

$$\hat{\beta}_{ls} = (k^T k)^{-1} k^T \tilde{y} = (H^T WH)^{-1} H^T W Y$$

The error cov. matrix (with $E[N] = E[u] = 0$) would be:

$$\text{Cov} \{ \hat{\beta}_{ls} \} = (H^T WH)^{-1} H^T W \text{Cov}(N) WH (H^T WH)^{-1}$$

(show if you don't see it).
14 Identifiability in the Additive Linear Model

1. Here we consider the previous additive model $Y = H\theta + N$ where the random vector $N$ has zero mean. Suppose that $H$ is $m \times n$ where $m \geq n > \text{rank}(H)$.

2. While the distribution of the observation $Y$ is centered on $H\theta$ and certainly depends upon $\theta$, different values of $\theta$ give the same value for $H\theta$ and hence the same distribution on the sample space.

3. Hence, while the observation $Y = y$ may give some information about $H\theta$ it cannot help decide between $\theta$ and $\theta'$ if $H\theta = H\theta'$. The parameter $\theta$ is said to be unidentifiable.

4. Can talk about equivalence classes of parameters or about re-parameterizing to avoid the problem. Both are legitimate but sometimes the “natural” parameterization leads to such a problem.

5. More of an annoyance than a deep issue.
15 The Gauss-Markov Theorem

15.1 Reprise of Least Squares

1. Here we consider the previous additive model \( Y = H\theta + N \) where the random vector \( N \) has zero mean. Suppose that \( H \) is \( m \times n \) where \( m \geq n \geq \text{rank}(H) \).

2. Intuitive justification for least-squares: Estimate the mean of a distribution by the parameter nearest to the observation made, i.e., given \( Y = y \) the mean estimate is 

\[
\hat{H}\theta \approx y
\]

3. If the parameter vector is identifiable (\( \text{rank}(H) = n \)) then the least-squares estimate of the parameter \( \theta \) is

\[
\hat{\theta}_{LS} = (H^T H)^{-1} H^T y.
\]

4. Under these assumptions the least-squares estimator is unbiased, i.e.,

\[
E_{\theta}\{\hat{\theta}_{LS}(Y)\} = \theta
\]

for all \( \theta \in \mathcal{R}^n \). Also, the covariance matrix of \( \hat{\theta}_{LS}(Y) \) is

\[
\text{Cov}_{\theta}\{\hat{\theta}_{LS}\} = E_{\theta}\{(\hat{\theta}_{LS} - \theta)(\hat{\theta}_{LS} - \theta)^T\}
\]

\[
= (H^T H)^{-1} H \text{Cov}(N) H^T (H^T H)^{-1}
\]

\[
= (H^T H)^{-1} H \Sigma N H^T (H^T H)^{-1}.
\]
15.2 Usually Least Squares is Not the Best We Can Do

1. Suppose that \( \hat{\theta}_a(Y) \) and \( \hat{\theta}_b(Y) \) are two estimators of the vector parameter \( \theta \).

2. Suppose both are unbiased, i.e., \( \theta = \mathbb{E}_{\theta}\{\hat{\theta}_a(Y)\} = \mathbb{E}_{\theta}\{\hat{\theta}_b(Y)\} \) for all \( \theta \).

3. Then we are interested in performance measures such as the covariance matrix

\[
\operatorname{Cov}_\theta\{\hat{\theta}_a(Y)\} = \mathbb{E}_\theta\{(\hat{\theta}_a(Y) - \theta)(\hat{\theta}_a(Y) - \theta)^T\}
\]

and the mean squared error (MSE)

\[
\epsilon^2_{ms}(\hat{\theta}) = \mathbb{E}_\theta\{(\hat{\theta}_a(Y) - \theta)^T(\hat{\theta}_a(Y) - \theta)\}
\]

\[
= \sum_{k=1}^{n} \mathbb{E}_\theta\{(\hat{\theta}_{a,k}(Y) - \theta_k)^2\}
\]

\[
= \text{Tr } \operatorname{Cov}_\theta\{\hat{\theta}_a(Y)\}
\]

4. Then we might say that \( \hat{\theta}_a(Y) \) is preferable to \( \hat{\theta}_b(Y) \) provided the following equivalent relationships hold:

(a) \( \operatorname{Cov}_\theta\{\hat{\theta}_b(Y)\} - \operatorname{Cov}_\theta\{\hat{\theta}_a(Y)\} \geq 0 \), i.e., the difference between the two covariance matrices is non-negative definite.

(b) \( \epsilon^2_{ms}(\hat{\theta}_b) \geq \epsilon^2_{ms}(\hat{\theta}_a) \).
(c) For any $n$-vector $c$

$$\text{Var}_\theta\{c^T\hat{\theta}_b\} \geq \text{Var}_\theta\{c^T\hat{\theta}_a\}.$$
We have looked at the error covariance matrices of weighted and unweighted least squares because these things provide a measure of estimator behavior and goodness.

It is therefore natural to try to find the estimator which minimizes error variances. One way to go about this is by looking at the mean squared error criterion:

The mean square error (MSE) of an estimator \( \hat{\theta} \) of a parameter vector \( \theta \) is

\[
E_{\text{MSE}} = E\{(\theta - \hat{\theta})(\theta - \hat{\theta})^T\} = \sum_{i=1}^{n} E\{\hat{\theta}_i^2\}
\]

Note that for an unbiased estimator \( E\{\hat{\theta}\} = \theta \) the MSE above is the sum of the variance of each component.

The reason for bringing this up now is to make the connection between the LS estimator and the estimator that minimizes MSE for the linear additive noise model

\[
Y = H\theta + N \quad H \text{ m x n and of rank = n} \quad E\{N\} = 0 \quad E\{NN^T\} = \Sigma_n
\]

Consider linear estimators \( \hat{\theta}(Y) = \hat{\theta} \) of the form

\( \hat{\theta} = K\hat{Y} \quad K \text{ m x n} \)

We want to require \( \hat{\theta} \) to be unbiased \( (\theta = E\{\hat{\theta}\}) \) which then requires

\[
\theta = E\{\hat{\theta}\} = E\{K\hat{Y}\} = KE\{H\theta + N\} = K\theta
\]

So in this context, the estimator is unbiased if and only if

\[
KH = I_n \quad \text{(in particular, } H \text{ must admit a left inverse } \iff m > n = \text{rank}(H))
\]

Then computing the estimation error vector

\[
\tilde{\theta} = \theta - \hat{\theta} = \theta - KY = \theta - K(H\theta + N) = -KN
\]

and the MSE of such an estimator is

\[
E_{\text{MSE}} = E\{\tilde{\theta}\tilde{\theta}^T\} = \text{Trace} E\{\tilde{\theta}\tilde{\theta}^T\} = \text{Trace} \Sigma_n K^T
\]

We have shown that the linear unbiased estimator that minimizes the MSE is found by determination of the matrix \( K \) (m x n)

\[
\text{minimizing} \quad \text{Trace} \Sigma_n K^T \quad \text{subject to} \quad KH = I_n
\]

We will discuss the solution to this problem presently. First however, we remark on the importance of the unbiased constraint which we have made.

Suppose that we had only posed the following

\[
\hat{\theta} = KY \quad \text{(linear, K nxm)}
\]

and pick \( K \) to minimize the MSE:

\[
E_{\text{MSE}} = E\{\tilde{\theta}\tilde{\theta}^T\}
\]
Consider
\[ \mathbf{E'} = (\mathbf{E-K})' \mathbf{(K-Y)} \]
\[ = [(\mathbf{K-KH})-\mathbf{KN}]' [\mathbf{K-K'E}] \mathbf{KN} \]

Multiplying out, taking \( \mathbb{E} \{ \cdot \} \) and using fact that \( \mathbf{N} \) is zero mean we have:
\[ \mathbb{E}_{15'} = \|\mathbf{E-K}\mathbf{N}\|^2 + \mathbb{E} \{ \|\mathbf{K}\mathbf{N}\|^2 \} \]
\[ = \|\mathbf{E-K}\mathbf{N}\|^2 + \text{Trace} \mathbf{K_zN} \mathbf{K'} \]  

The presence of this term would mean that the optimal \( \mathbf{K} \) and hence optimal \( \mathbf{E} \) is a function of the true parameter \( \mathbf{E} \).

Thus we go to the modified or constrained problem of requiring \( \mathbf{KH} = \mathbf{I} \Rightarrow \) unbiased.

The Gauss-Markov Theorem

Suppose that \( \mathbf{Y} = \mathbf{H} + \mathbf{N} \) where \( \mathbf{H} \) is \( m \times n \) and of rank \( k \). \( \mathbf{E} \mathbf{N} = \mathbf{0}, \mathbf{E} \mathbf{N}' \mathbf{N}' = \mathbf{A} \) then the linear minimum MSE unbiased estimate of \( \mathbf{E} \) is
\[ \hat{\mathbf{E}}_{15} = (\mathbf{H}' \mathbf{A}' \mathbf{H})^{-1} \mathbf{H}' \mathbf{A}' \mathbf{Y} \]
with corresponding error covariance:
\[ \mathbb{E} \{ (\hat{\mathbf{E}}_{15} - \mathbf{E})(\hat{\mathbf{E}}_{15} - \mathbf{E})' \} = (\mathbf{H}' \mathbf{A}' \mathbf{H})^{-1} \]
References


