Approximation: Linear Least Squares

Math 131: Numerical Analysis

J.C. Meza

March 19, 2024

Section 1

Introduction

Recall - Data Fitting

Suppose we are given a set of data points

$$\{(x_i, y_i)\}_{i=0}^{n}.$$

The points x_i are sometimes called the ***node points*** or simply just ***nodes***.

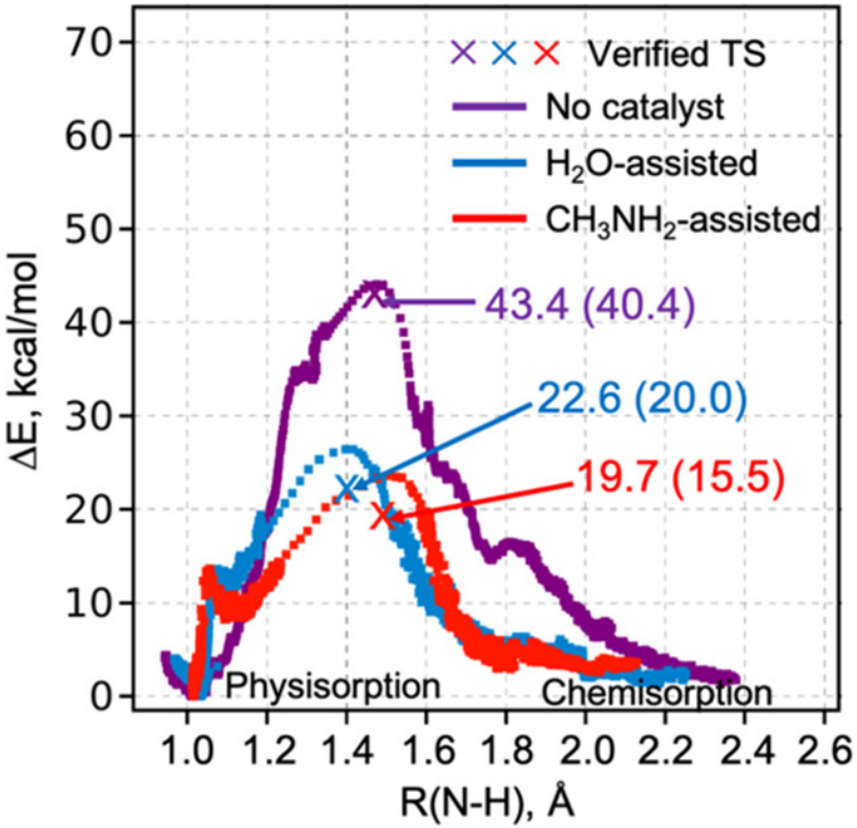
Goal

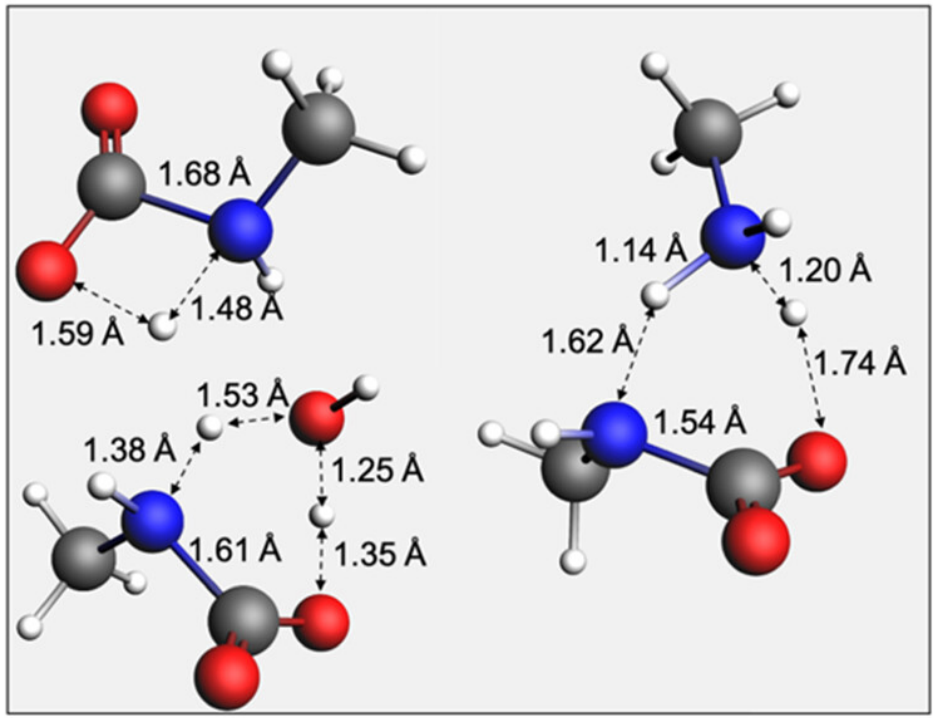
Find a function $v(x)$ that "fits" the data in some yet to be determined way.

Before:

- Considered the case where the approximation interpolated the data.
- What about situations where the data is known to have some error (experimental, observational, etc.)

Real Data





https://pubs.acs.org/doi/full/10.1021/acs.jpcc.3c07183

Real Data

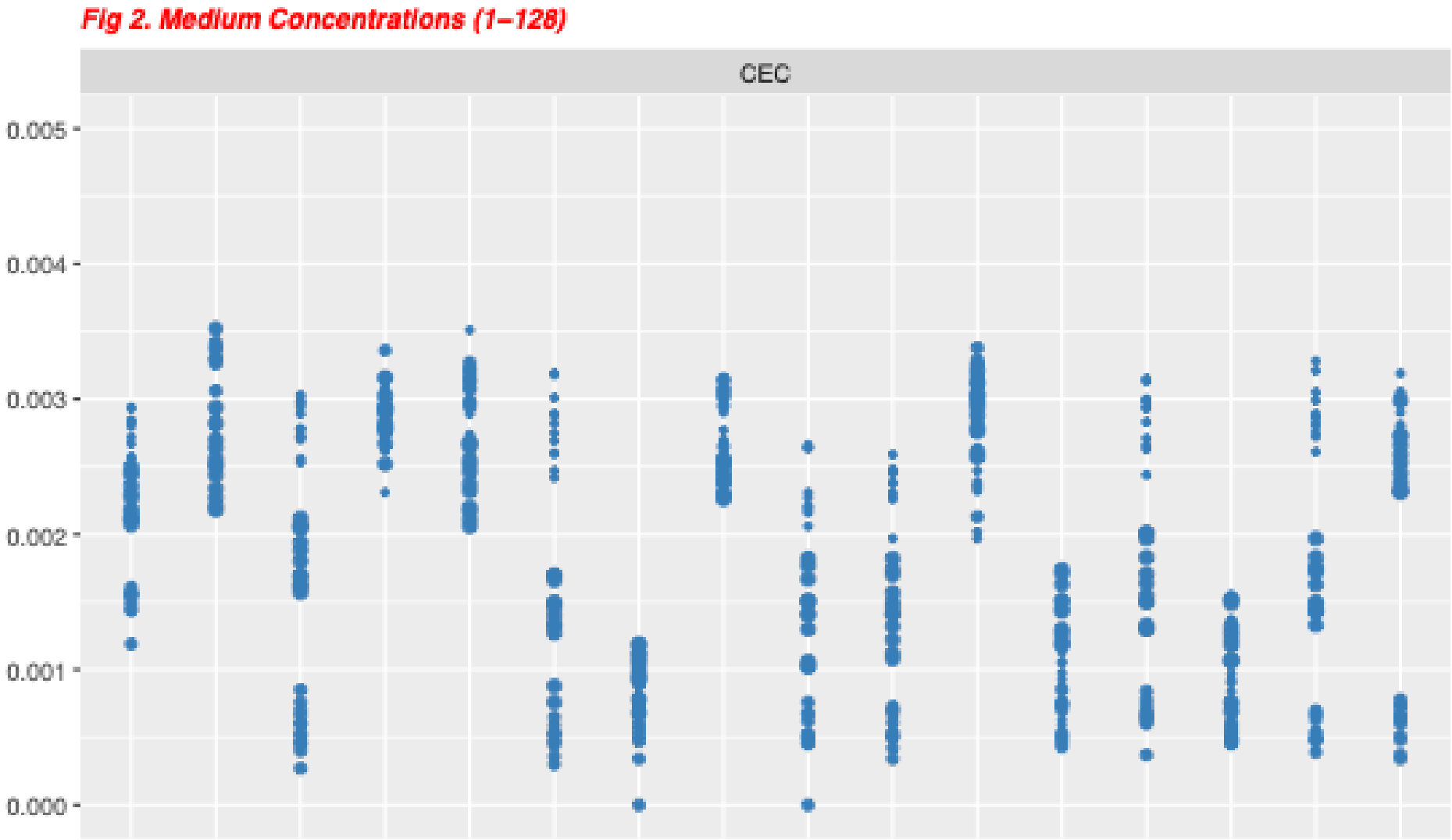


Figure 1: Growth Rate vs. Antibiotics

Section 2

Linear Algebra Review: Norms

Review (Vector Norms)

Consider $x \in \mathbb{R}^n$

By the l_2-*norm* we mean the standard Euclidean length:

$$||x||_2 = \sqrt{x^T x}$$

By the l_∞-*norm* we mean:

$$||x||_\infty = \max_{1 \leq i \leq n} |x_i|$$

By the l_1-*norm* we mean:

$$||x||_1 = \sum_{i=1}^{n} |x_i|$$

Example

Suppose we want to compute the error between 2 vectors, x and y

$$x = \begin{bmatrix} 1.0 \\ 2.0 \\ 3.0 \end{bmatrix} \qquad y = \begin{bmatrix} 1.1 \\ 2.2 \\ 10.0 \end{bmatrix} \qquad z = y - x = \begin{bmatrix} 0.1 \\ 0.2 \\ 7.0 \end{bmatrix}$$

$$\begin{aligned} ||x||_1 &= 0.1 + 0.2 + 7.0 = 7.3 \\ ||x||_2 &= \sqrt{0.1^2 + 0.2^2 + 97^2} = 7.0036 \\ ||x||_\infty &= 7 \end{aligned}$$

Note

All norms are the same order of magnitude in finite-dimensional spaces.

Review (Matrix Norms)

Consider $A \in \mathbb{R}^{m \times n}$. We can define an ***induced norm*** associated with each of the vector norms: $p = 1, 2, \infty$.

$$\begin{aligned} ||A||_p &= \max_{x \neq 0} \frac{||Ax||_p}{||x||_p} \\ &= \max_{||x||_p = 1} ||Ax||_p \end{aligned}$$

The matrix norms can be shown to satisfy the following 4 properties:

1. $||A|| \geq 0; ||A|| = 0 \iff A = 0;$
2. $||\alpha A|| = |\alpha| \; ||A|| \quad \forall \alpha \in \mathbb{R};$
3. $||A + B|| \leq ||A|| + ||B|| \quad \forall A, B \in \mathbb{R}^{m \times n};$ ***triangle inequality***
4. $||A \cdot B|| \leq ||A|| \cdot ||B|| \quad \forall A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times k};$ ***consistency***

Matrix Norms (cont.)

This leads to similar formulas for the case of a matrix.

By the *infinity-norm* of a matrix we mean:

$$||A||_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|$$

By the 1-*norm* of a matrix we mean:

$$||A||_1 = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|$$

The $2-$*norm* of a matrix is more complicated:

$$||A||_2 = \sqrt{\rho(A^T A)} = \text{largest singular value of A}$$

Section 3

Least Squares

Back to the Least Squares problem

Idea

Minimize the errors between the data and the model using one of the above norms, for example the l_2-norm.

1. First construct a model as before, e.g.

$$f(x) = c_0\phi_0(x) + c_1\phi_1(x) + \ldots + c_n\phi_n(x)$$

2. Minimize the difference between the function and the data points, e.g.

$$\min_{c_0,\ldots,c_n} F(c_0, \ldots, c_n) = \sum_{i=0}^{m} (f(x_i) - y_i)^2$$

Note

This is usually referred to as minimizing the ***sum of squares***

Choice of basis functions

Basis function vary widely and depend on the problem to be solved. Examples:

1. $\phi_j(x) = x^j$
2. $\phi_j(x) = e^{\beta_j x}$, β_j are known
3. $\phi_j(x) = \sin(\beta_j x)$, β_j are known
4. Piecewise-polynomials or even mixed basis functions

$$\phi_0(x) = x^3, \quad \phi_1(x) = 1/x, \quad \phi_2(x) = \ln(x)$$
$$f(x) = c_0 x^3 + c_1/x + c_2 \ln(x)$$

Remark

The important thing to note is that the coefficients appear ***linearly*** in the equation.

General Case

This leads us to set up the minimization problem as:

$$\begin{aligned}\min_{c_0,\dots,c_n} F(c_0,\dots,c_n) &= \sum_{k=0}^{m}(f(x_k)-y_k)^2 \\ &= \sum_{k=0}^{m}\Big(c_0\phi_0(x_k)+\dots+c_n\phi_n(x_k)-y_k\Big)^2\end{aligned}$$

If we notice that this is really just a quadratic function, then it makes sense to take the derivative and set it equal to zero. But now because we're not in 1D, we need to take all of the partial derivatives, i.e.

$$\frac{\partial}{\partial c_j}F(c_0,\dots,c_n)=0, \quad j=0,\dots,n.$$

General Case (cont.)

Let's first consider one of the partial derivatives. Note that

$$\frac{\partial}{\partial c_j} F(c_0, \dots, c_n) = 2 \sum_{k=0}^{m} \Big(c_0 \phi_0(x_k) + \dots + c_n \phi_n(x_k) - y_k \Big) \cdot \phi_j(x_k)$$

For $j = 0$ this looks like

$$\sum_{k=0}^{m} c_0 \phi_0(x_k)\phi_0(x_k) + \dots + c_n \phi_n(x_k)\phi_0(x_k) = \sum_{k=0}^{m} y_k \phi_0(x_k)$$

For $j = n$ this looks like

$$\sum_{k=0}^{m} c_0 \phi_0(x_k)\phi_n(x_k) + \dots + c_n \phi_n(x_k)\phi_n(x_k) = \sum_{k=0}^{m} y_k \phi_n(x_k)$$

Matrix notation

In matrix notation we can write this as:

$$\begin{bmatrix} \sum \phi_0(x_k)^2 & \dots & \sum \phi_n(x_k)\phi_0(x_k) \\ \vdots & \vdots & \vdots \\ \sum \phi_0(x_k)\phi_n(x_k) & \dots & \sum \phi_n(x_k)^2 \end{bmatrix} \begin{bmatrix} c_0 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \sum y_i\phi_0(x_k) \\ \vdots \\ \sum y_i\phi_n(x_k) \end{bmatrix} \quad (1)$$

Note: all summations run from $k = 0, \dots, m$

or more compactly as:

$$Bc = y$$

General Term

In fact, we can write down the general term for any of the entries in the matrix as:

$$B_{ij} = \sum_{k=0}^{m} \phi_i(x_k)\phi_j(x_k)$$

Properties of matrix

There are several important properties of the B matrix: 1) symmetric, 2) positive definite, 3) good chance for ill-conditioning when m, n are large.

Example for n=2

Suppose we have only 2 basis functions and we choose them as $\phi_0 = 1, \phi_1 = x$

Then using Equation 1 we can show that the equations we need to solve are:

$$\begin{bmatrix} \sum 1 & \sum x_k \\ \sum x_k & \sum x_k^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} \sum y_k \\ \sum y_k x_k \end{bmatrix}$$

Note: all summations run from $k = 0, \ldots, m$

Section 4

Linear Least Squares - Normal Equations

Problem Statement

We can state the problem of fitting data by the following:

$$\min_x \frac{1}{2}||b - Ax||^2$$

where $A \in \mathbb{R}^{mxn}, x \in \mathbb{R}^n, b \in \mathbb{R}^m,\ m \geq n.$

The quantity $r = b - Ax$ is called the ***residual.***

Minimization problem

Using the definition of the residual we can restate the minimization problem:

$$\min_x \frac{1}{2}||b - Ax||^2 = \min_x \Psi(x)$$

where

$$\psi(x) = \frac{1}{2}||b - Ax||^2 = \frac{1}{2}||r||^2$$

Geometrically

It's much easier to see the conditions for a minimum in a figure. Note that the residual is orthogonal to the column space of A.

Normal equations

From our picture we have:

$$A^T r = 0$$

Using the definition of r and rearranging the equation:

$$A^T r = A^T(b - Ax) = A^T b - A^T Ax = 0$$

Leading to what are called the:

Normal equations

$$A^T Ax = A^T b$$

Theory

When do we have a solution?

- Does a minimizer exist, i.e. a solution to the normal equations?
- If so, is the solution a global miminizer?
- Is the problem well-conditioned? Ill-conditioned?

LSQ uniqueness

Unique solution to normal equations

If A has full column rank then $A^T A$ is ***symmetric positive definite*** (s.p.d) and the least squares problem

$$\min_x \frac{1}{2}||b - Ax||^2$$

has a unique solution that satisifes the normal equations.

Solution to normal equations

First write down the normal equations

$$A^T A x = A^T b$$

Can write the solution to the normal equations as $x = (A^T A)^{-1} A^T b$

Pseudo-inverse

The term

$$A^\dagger = (A^T A)^{-1} A^T$$

is called the ***pseudo-inverse***.

Think of the pseudo-inverse as a generalization of the standard matrix inverse for the case where the matrix A is not square

Algorithm: Solving the Normal Equations

1. Form $B = A^T A$ and $y = A^T b$
2. Compute the Cholesky factors of B ($B = GG^T$).
3. Solve the lower triangular system $Gz = y$ for z
4. Solve the upper triangular system $G^T x = z$ for x

Section 5

Example

Data Fitting Example

- Let's consider fitting a set of m data points, (t_i, b_i) by a straight line:

$$v(t) = x_1 + x_2 t$$

- Unlike interpolation, this time we will only require an approximate fit:

$$v(t_i) \approx b_i, \quad i = 1, \ldots, m.$$

where we want to:

$$\min \ ||b - Ax||^2$$

$$A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix}, \ x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \ b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix},$$

Note: x is the unknown and b corresponds to the function values.

Exercise

Construct least squares approximation to the following data:

	$i = 1$	$i = 1$	$i = 3$
t_i	0.0	1.0	2.0
b_i	0.1	0.9	2.0

Solution

Step 1. Setup

$$A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1.0 \\ 1 & 2.0 \end{bmatrix}, \qquad A^T = \begin{bmatrix} 1 & 1 & 1 \\ t_1 & t_2 & t_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1.0 & 2.0 \end{bmatrix}$$

$$B = A^T A = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix} \qquad y = A^T b = \begin{bmatrix} 3 \\ 4.9 \end{bmatrix}$$

Solution (cont.)

Step 2. Solve for x using $Bx = y$:

$$\begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix} x = \begin{bmatrix} 3 \\ 4.9 \end{bmatrix}$$

$$x = \begin{bmatrix} 0.5 \\ 0.95 \end{bmatrix}$$

Tip

Note: One can use any standard linear solver. In practice, you should use a linear solver specific to symmetric matrices (Cholesky decomposition).

Notes on solution

- Unlike interpolation, the linear approximation only gets close to the data points
- In fact, it's easy to see that the residual is nonzero

$$r = b - Ax = \begin{bmatrix} 0.05 \\ -0.10 \\ 0.05 \end{bmatrix}$$

and

$$||r||_2 = 0.1225$$

Practical Tips

- This approach is easy to set up and solve
- The normal equations generate a s.p.d matrix
- This implies that B is nonsingular and hence the linear system has a unique solution.
- Good linear solvers available for solving s.p.d. matrix
- Unfortunately, this problem may be ill-conditioned, so care must be taken. Can show that the condition number for the $A^T A$ matrix is the square of the condition number of A.

Using other norms

- We've restricted ourselves to the l_2 norm, which is common
- Other norms can be used, for example l_1, which are popular for certain problems.
- One other norm you might see is the ***infinity*** norm, l_∞.

Section 6

Summary

Summary -

- Introduced the general linear least squares problem
- The normal equations can be used to find a solution to the least squares problem.
- Applications of least squares abound in all areas of science and engineering.

Section 7

Supplemental Materials (Optional)

Alternate derivation of normal equations

We can state the problem of fitting data by the following:

$$\min_x \frac{1}{2}||b - Ax||^2$$

where $A \in \mathbb{R}^{mxn}, x \in \mathbb{R}^n, b \in \mathbb{R}^m,\ m \geq n.$

The quantity $r = b - Ax$ is called the ***residual.***

Minimization problem

Using the definition of the residual we can restate the minimization problem:

$$\min_x \frac{1}{2}||b - Ax||^2 = \min_x \Psi(x)$$

where

$$\psi(x) = \frac{1}{2}||b - Ax||^2 = \frac{1}{2}||r||^2$$

Necessary conditions

- As in the 1D case, our approach is to take the derivative and set it equal to 0.
- As before, because x is a vector, we need to take the derivative of $\psi(x)$ with respect to each of the components of x.

$$\frac{\partial}{\partial x_k}\psi(x) = 0, \quad k = 1, \dots, n.$$

Taking derivatives

If

$$\psi(x) = \frac{1}{2}\sum_{i=1}^{m}(b_i - Ax_i)^2$$

Then (as before):

$$\frac{\partial}{\partial x_k}\psi(x) = \sum_{i=1}^{m}\left[\left(b_i - \sum_{j=1}^{n} a_{ij}x_j\right)(-a_{ik})\right] = 0, \quad k = 1, \dots, n$$

where the last part can be rewritten as:

$$\sum_{i=1}^{m} a_{ik}\sum_{j=1}^{n} a_{ij}x_j = \sum_{i=1}^{m} a_{ik}b_i, \quad k = 1, \dots, n$$

or

$$A^T Ax = A^T b.$$