# Detailed Derivation of the Least Squares Solution: From Residual Gradients to the Normal Equations

Addendum to the Lecture Notes of Lecture 7

## 1 Problem Formulation and Matrix Representation

We aim to fit a straight line modeled by the linear function $y = ax + b$ through a dataset containing exactly three distinct experimental measurements: $(x_1, y_1)$, $(x_2, y_2)$, and $(x_3, y_3)$.

Ideally, if all data points were perfectly collinear, they would satisfy the following overdetermined linear system exactly:

$$\begin{aligned} ax_1 + b &= y_1, \\ ax_2 + b &= y_2, \\ ax_3 + b &= y_3. \end{aligned}$$

We can compress this system into a clean matrix-vector formulation $X\boldsymbol{\beta} \approx \mathbf{y}$ by defining a $3 \times 2$ design matrix $X$, a $2 \times 1$ parameter vector $\boldsymbol{\beta}$, and a $3 \times 1$ observation vector $\mathbf{y}$:

$$X = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}. \tag{1}$$

## 2 The Objective Function: Sum of Squared Residuals

Because experimental measurements inherently contain noise, an exact solution to $X\boldsymbol{\beta} = \mathbf{y}$ typically does not exist. Instead, we define a residual vector $\mathbf{r}(\boldsymbol{\beta}) \in \mathbb{R}^3$ which captures the discrepancy between the observed values and our model's predictions:

$$\mathbf{r}(\boldsymbol{\beta}) = \mathbf{y} - X\boldsymbol{\beta} = \begin{bmatrix} y_1 - (ax_1 + b) \\ y_2 - (ax_2 + b) \\ y_3 - (ax_3 + b) \end{bmatrix}. \tag{2}$$

To find the optimal parameter vector $\hat{\boldsymbol{\beta}} = [\hat{a}, \hat{b}]^T$, the method of Ordinary Least Squares minimizes the scalar objective function $S(\boldsymbol{\beta})$, which represents the sum of the squared residuals ($L_2$-norm):

$$S(\boldsymbol{\beta}) = \|\mathbf{r}(\boldsymbol{\beta})\|^2 = \mathbf{r}(\boldsymbol{\beta})^T \mathbf{r}(\boldsymbol{\beta}) = \sum_{i=1}^{3} (y_i - (ax_i + b))^2. \tag{3}$$

Using the algebraic properties of matrix transposition, we expand the scalar function $S(\boldsymbol{\beta})$ as follows:

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\mathbf{y} - X\boldsymbol{\beta})^T (\mathbf{y} - X\boldsymbol{\beta}) \\ &= (\mathbf{y}^T - \boldsymbol{\beta}^T X^T)(\mathbf{y} - X\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\boldsymbol{\beta} - \boldsymbol{\beta}^T X^T \mathbf{y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}. \end{aligned}$$

Because each term in this expansion evaluates to a scalar, and a scalar is equal to its own transpose, we can simplify the inner terms. Specifically, notice that $(\mathbf{y}^T X\boldsymbol{\beta})^T = \boldsymbol{\beta}^T X^T \mathbf{y}$. Grouping these matching terms together yields our final expanded objective function:

$$S(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T X^T \mathbf{y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}. \tag{4}$$

## 3 Detailed Gradient Derivation via Matrix Calculus

To find the minimum of $S(\boldsymbol{\beta})$, we apply the necessary first-order optimality condition from multivariable calculus. We compute the gradient of $S$ with respect to the parameter vector $\boldsymbol{\beta}$ and set it equal to the zero vector:

$$\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial S}{\partial a} \\ \frac{\partial S}{\partial b} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{5}$$

Let us differentiate Eq. (4) term by term using fundamental matrix calculus rules:

1. **First Term ($\mathbf{y}^T\mathbf{y}$):** This term contains no parameters from $\boldsymbol{\beta}$. It acts as a constant, meaning its derivative is zero:
$$\nabla_{\boldsymbol{\beta}}(\mathbf{y}^T\mathbf{y}) = \mathbf{0}.$$

2. **Second Term ($-2\boldsymbol{\beta}^T X^T\mathbf{y}$):** Using the linear identity $\nabla_{\mathbf{z}}(\mathbf{z}^T\mathbf{v}) = \mathbf{v}$, we evaluate the derivative with respect to $\boldsymbol{\beta}$ where $\mathbf{v} = X^T\mathbf{y}$:
$$\nabla_{\boldsymbol{\beta}}(-2\boldsymbol{\beta}^T X^T\mathbf{y}) = -2X^T\mathbf{y}.$$

3. **Third Term ($\boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$):** This represents a quadratic form $\boldsymbol{\beta}^T M\boldsymbol{\beta}$ where $M = X^T X$. Because $M$ is a symmetric matrix ($M^T = (X^T X)^T = X^T X = M$), we apply the quadratic matrix identity $\nabla_{\mathbf{z}}(\mathbf{z}^T M\mathbf{z}) = 2M\mathbf{z}$:
$$\nabla_{\boldsymbol{\beta}}(\boldsymbol{\beta}^T X^T X\boldsymbol{\beta}) = 2X^T X\boldsymbol{\beta}.$$

Combining these three components together, the total gradient of our least squares cost function is:

$$\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = -2X^T\mathbf{y} + 2X^T X\boldsymbol{\beta}. \tag{6}$$

## 4 The Normal Equations and the Final Solution

Setting the total gradient vector Eq. (6) to zero to minimize the total squared error yields:

$$-2X^T\mathbf{y} + 2X^T X\boldsymbol{\beta} = \mathbf{0} \implies 2X^T X\boldsymbol{\beta} = 2X^T\mathbf{y}.$$

Dividing both sides by 2 produces the foundational **Normal Equations**:

$$(X^T X)\boldsymbol{\beta} = X^T\mathbf{y}. \tag{7}$$

Provided that the columns of the design matrix $X$ are linearly independent (which holds true as long as we have at least two distinct $x_i$ values), the square matrix $X^T X \in \mathbb{R}^{2\times 2}$ is invertible. We isolate the optimal estimator $\hat{\boldsymbol{\beta}}$ by multiplying by the inverse matrix:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (X^T X)^{-1} X^T\mathbf{y}. \tag{8}$$

### 4.1 Explicit Algebraic Structure of the Matrices

To give a complete overview, let us evaluate the explicit components of these matrices using our specific $3 \times 2$ formulation from Eq. (1):

$$X^T X = \begin{bmatrix} x_1 & x_2 & x_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{bmatrix} = \begin{bmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & 3 \end{bmatrix}.$$

Similarly, the right-hand vector $X^T\mathbf{y} \in \mathbb{R}^2$ structures as:

$$X^T\mathbf{y} = \begin{bmatrix} x_1 & x_2 & x_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \sum x_i y_i \\ \sum y_i \end{bmatrix}.$$

Using the analytical formula for the inverse of a $2 \times 2$ matrix, $(X^TX)^{-1}$ expands directly to:

$$(X^TX)^{-1} = \frac{1}{3\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} 3 & -\sum x_i \\ -\sum x_i & \sum x_i^2 \end{bmatrix}.$$

Substituting these expanded configurations into Eq. (8) provides the exact analytical linear fit parameters for any three given data points.

# 5 Numerical Example with Concrete Measurements

To demonstrate the practical application of the derived formulas, let us compute the optimal linear fit $y = ax + b$ for exactly three concrete experimental data points:

$$(x_1, y_1) = (1, 2), \quad (x_2, y_2) = (2, 3), \quad (x_3, y_3) = (3, 5).$$

## 5.1 Constructing the System Matrices

Following the definitions established in Section 1, we populate the $3 \times 2$ design matrix $X$ and the $3 \times 1$ observation vector $\mathbf{y}$ with our numerical values:

$$X = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix}.$$

## 5.2 Step-by-Step Matrix Operations

We now execute the sequential computations required by the Ordinary Least Squares solution $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$.

1. **Compute the Matrix Product $X^TX$:**

$$X^TX = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 1^2 + 2^2 + 3^2 & 1 + 2 + 3 \\ 1 + 2 + 3 & 1 + 1 + 1 \end{bmatrix} = \begin{bmatrix} 14 & 6 \\ 6 & 3 \end{bmatrix}.$$

2. **Compute the Right-Hand Vector $X^T\mathbf{y}$:**

$$X^T\mathbf{y} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} (1 \cdot 2) + (2 \cdot 3) + (3 \cdot 5) \\ 2 + 3 + 5 \end{bmatrix} = \begin{bmatrix} 2 + 6 + 15 \\ 10 \end{bmatrix} = \begin{bmatrix} 23 \\ 10 \end{bmatrix}.$$

3. **Invert the $2 \times 2$ Matrix $X^TX$:** The determinant of $X^TX$ is calculated as $\det(X^TX) = (14 \cdot 3) - (6 \cdot 6) = 42 - 36 = 6$. Utilizing the analytical inversion formula, we obtain:

$$(X^TX)^{-1} = \frac{1}{6} \begin{bmatrix} 3 & -6 \\ -6 & 14 \end{bmatrix} = \begin{bmatrix} \frac{3}{6} & -\frac{6}{6} \\ -\frac{6}{6} & \frac{14}{6} \end{bmatrix} = \begin{bmatrix} 0.5 & -1 \\ -1 & \frac{7}{3} \end{bmatrix}.$$

### 5.3 Calculating the Optimal Parameters

Finally, we multiply the inverted matrix $(X^TX)^{-1}$ by the vector $X^T\mathbf{y}$ to isolate our parameter vector $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned}\hat{\boldsymbol{\beta}} = \begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix} &= (X^TX)^{-1}X^T\mathbf{y}\\ &= \begin{bmatrix}\frac{1}{2} & -1\\ -1 & \frac{7}{3}\end{bmatrix}\begin{bmatrix}23\\ 10\end{bmatrix}\\ &= \begin{bmatrix}\left(\frac{1}{2}\cdot 23\right) + (-1\cdot 10)\\ (-1\cdot 23) + \left(\frac{7}{3}\cdot 10\right)\end{bmatrix}\\ &= \begin{bmatrix}11.5 - 10\\ -23 + \frac{70}{3}\end{bmatrix} = \begin{bmatrix}1.5\\ -\frac{69}{3} + \frac{70}{3}\end{bmatrix} = \begin{bmatrix}\frac{3}{2}\\ \frac{1}{3}\end{bmatrix}.\end{aligned}$$

The optimal parameters are $\hat{a} = 1.5$ (the slope) and $\hat{b} = \frac{1}{3} \approx 0.333$ (the y-intercept). The resulting best-fitting linear model for this dataset is:

$$y = 1.5x + \frac{1}{3}. \tag{9}$$

This concrete numerical breakdown allows students to verify that the abstract matrix algebra precisely evaluates to the correct scaling parameters.

## 6 Algebraic Derivation: Term-by-Term Expansion with Concrete Numbers

To anchor the abstract matrix derivation, we now bypass all matrix algebra, transposes, and vectors. We will write out the Ordinary Least Squares objective function $S(a, b)$ directly as a scalar algebraic sum using our concrete experimental measurements:

$$(x_1, y_1) = (1, 2), \quad (x_2, y_2) = (2, 3), \quad (x_3, y_3) = (3, 5).$$

Our goal is to fit the linear model $y = ax + b$ by minimizing the sum of the squared individual residuals.

### 6.1 Step 1: Fully Expanding the Objective Function

The objective function $S(a, b)$ is defined as the sum of the squared differences between the observed $y_i$ and the predicted values $ax_i + b$:

$$\begin{aligned}S(a,b) &= \sum_{i=1}^{3}\left(y_i - (ax_i + b)\right)^2\\ &= (2 - (a(1) + b))^2 + (3 - (a(2) + b))^2 + (5 - (a(3) + b))^2\\ &= (2 - a - b)^2 + (3 - 2a - b)^2 + (5 - 3a - b)^2.\end{aligned}$$

We now expand each of the three quadratic terms individually using the algebraic identity $(A - B - C)^2 = A^2 + B^2 + C^2 - 2AB - 2AC + 2BC$:

1. **First Term:**
$$(2 - a - b)^2 = 4 + a^2 + b^2 - 4a - 4b + 2ab$$

2. **Second Term:**
$$(3 - 2a - b)^2 = 9 + 4a^2 + b^2 - 12a - 6b + 4ab$$

3. **Third Term:**
$$(5 - 3a - b)^2 = 25 + 9a^2 + b^2 - 30a - 10b + 6ab$$

Summing these three equations horizontally by grouping like terms ($a^2$, $b^2$, $ab$, $a$, $b$, and constants), we find the fully expanded scalar objective function:

$$\begin{aligned} S(a,b) &= (1+4+9)a^2 + (1+1+1)b^2 + (2+4+6)ab - (4+12+30)a - (4+6+10)b + (4+9+25) \\ &= 14a^2 + 3b^2 + 12ab - 46a - 20b + 38. \end{aligned}$$

### 6.2 Step 2: Term-by-Term Differentiation (Computing the Gradient)

To find the minimum, we take the partial derivatives of $S(a,b)$ with respect to our two parameters, $a$ and $b$, and set them equal to zero. This defines the explicit components of our gradient vector $\nabla S = [0,0]^T$:

1. **Partial Derivative with respect to $a$:** Treating $b$ as a constant and differentiating $S(a,b) = 14a^2 + 3b^2 + 12ab - 46a - 20b + 38$:
$$\frac{\partial S}{\partial a} = \frac{d}{da}(14a^2) + 0 + \frac{d}{da}(12ab) - \frac{d}{da}(46a) - 0 + 0 = 28a + 12b - 46. \tag{10}$$

2. **Partial Derivative with respect to $b$:** Treating $a$ as a constant and differentiating the same function:
$$\frac{\partial S}{\partial b} = 0 + \frac{d}{db}(3b^2) + \frac{d}{db}(12ab) - 0 - \frac{d}{db}(20b) + 0 = 6b + 12a - 20. \tag{11}$$

### 6.3 Step 3: Setting up and Solving the Linear System

Setting both partial derivatives to zero to satisfy the first-order optimality condition yields a system of two linear equations with two unknowns:

$$\begin{aligned} 28a + 12b &= 46, \\ 12a + 6b &= 20. \end{aligned}$$

To isolate $a$ and $b$, we can use the method of elimination. We multiply the second equation entirely by 2:
$$2 \cdot (12a + 6b = 20) \implies 24a + 12b = 40.$$

Now, we subtract this new equation from the first equation to eliminate the parameter $b$:

$$\begin{aligned} (28a + 12b) - (24a + 12b) &= 46 - 40 \\ 4a &= 6 \\ \hat{a} &= \frac{6}{4} = 1.5. \end{aligned}$$

With the optimal slope $\hat{a} = 1.5$ determined, we substitute this value back into the second original equation ($12a + 6b = 20$) to find the intercept $b$:

$$\begin{aligned} 12(1.5) + 6b &= 20 \\ 18 + 6b &= 20 \\ 6b &= 2 \\ \hat{b} &= \frac{2}{6} = \frac{1}{3}. \end{aligned}$$

### 6.4 Conclusion: Convergence of Methods

The scalar algebraic calculation yields the exact same optimal linear model:

$$y = 1.5x + \frac{1}{3}.$$

Notice the beautiful symmetry: the coefficients of our system equations (28, 12, 6, 3) correspond exactly to the entries of the matrix product $2X^TX$ and $2X^T\mathbf{y}$ from the previous matrix-based numerical chapter. This guarantees to the students that matrix calculus is not a different mathematical theory, but a powerful, structured shorthand for handling multi-variable calculus efficiently.

## 7 Bridging the Gap: Mapping Algebra Back to Matrix Notation

Now that we have computed the optimal parameters using both abstract matrix calculus and explicit scalar algebra, let us directly map the scalar results back to the matrix structures. This comparison explicitly highlights *why* we use matrix notation in computational sciences.

### 7.1 Mapping the Gradient Components

In our scalar derivation (Section 5), setting the partial derivatives to zero yielded the following two linear equations:

$$\begin{aligned} 28a + 12b &= 46, \\ 12a + 6b &= 20. \end{aligned}$$

If we pull the constants back to the left-hand side, this system represents the two components of the exact scalar gradient vector $\nabla S(a, b) = [0, 0]^T$:

$$\begin{bmatrix} \frac{\partial S}{\partial a} \\ \frac{\partial S}{\partial b} \end{bmatrix} = \begin{bmatrix} 28a + 12b - 46 \\ 12a + 6b - 20 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{12}$$

Now, let us revisit our abstract matrix gradient formula derived in Section 3, given by Eq. (6):

$$\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = 2X^TX\boldsymbol{\beta} - 2X^T\mathbf{y}.$$

Substituting the concrete numerical components of $X^TX$ and $X^T\mathbf{y}$ that we computed in Section 4 into this abstract formula, we get:

$$\begin{aligned} \nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) &= 2 \begin{bmatrix} 14 & 6 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} - 2 \begin{bmatrix} 23 \\ 10 \end{bmatrix} \\ &= \begin{bmatrix} 2(14a + 6b) \\ 2(6a + 3b) \end{bmatrix} - \begin{bmatrix} 46 \\ 20 \end{bmatrix} \\ &= \begin{bmatrix} 28a + 12b - 46 \\ 12a + 6b - 20 \end{bmatrix}. \end{aligned}$$

Notice that this matrix expression is **identically equal** to the vector of scalar partial derivatives in Eq. (12). Matrix multiplication automatically handles the collection of matching algebraic terms across all data points simultaneously.