Deriving the Least Squares Solution

Supplementary Material for Supervised Learning

Daniel E. Acuna
Associate Professor, University of Colorado Boulder

Ordinary Least Squares: Mathematical Derivation

In this supplementary material, we will:

- Develop the full mathematical derivation of OLS
- Use calculus to find the parameters that minimize squared error
- Express the solution in matrix form
- Derive the closed-form expressions for optimal parameters

The Least Squares Problem

We begin with our linear model:

$$\hat{y_i} = \beta_0 + \beta_1 x_i$$

Our goal is to find the values of β_0 and β_1 that minimize the sum of squared errors:

$$\sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Step 1: Define the Loss Function

We define a loss function $L(\beta_0, \beta_1)$ representing the sum of squared errors:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Step 2: Find Critical Points by Taking Partial Derivatives

For the optimal values of β_0 and β_1, the partial derivatives must equal zero:

$$\frac{\partial L}{\partial \beta_0} = 0 \quad \text{and} \quad \frac{\partial L}{\partial \beta_1} = 0$$

Step 3: Calculate the Gradients

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

Step 4: Simplify the Equations

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

Step 5: Rewrite as Normal Equations

Expanding the first equation:

$$\sum_{i=1}^{n} y_i - \beta_0 \sum_{i=1}^{n} 1 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

Which gives us:

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

Similarly for the second equation:

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Step 5.5: Matrix Notation for Linear Regression

Let's see how we can represent our regression problem using matrices. First, we define:

- **Design matrix $\mathbf{X}$**: A matrix with one row per data point, where the first column is all 1's (for the intercept) and the second column contains the x_i values
- **Parameter vector β**: Contains the regression coefficients $[\beta_0, \beta_1]^T$
- **Response vector $\mathbf{y}$**: Contains all the observed y_i values

For example, with n data points:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Step 5.6: Computing $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ - Part 1

Now, let's examine the matrix product $\mathbf{X}^{\mathrm{T}}\mathbf{X}$:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Step 5.6: Computing $\mathbf{X}^\mathrm{T}\mathbf{X}$ - Part 2

Computing this matrix multiplication:

$$\mathbf{X}^\mathrm{T}\mathbf{X} = \begin{bmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

Step 5.7: Computing $\mathbf{X}^\mathbf{T}\mathbf{y}$ - Part 1

Similarly, let's compute $\mathbf{X}^\mathbf{T}\mathbf{y}$:

$$\mathbf{X}^\mathbf{T}\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Step 5.7: Computing $\mathbf{X}^{\mathrm{T}}\mathbf{y}$ - Part 2

This gives us:

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

Step 5.8: Matrix Form of the Normal Equations - Part 1

Recall our normal equations:

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

We can write these in matrix form as:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

Step 5.8: Matrix Form of the Normal Equations - Part 2

Which is precisely:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Step 6: Express in Matrix Form

We can write this system as $\mathbf{X}^{\mathbf{T}}\mathbf{X}\beta = \mathbf{X}^{\mathbf{T}}\mathbf{y}$, where:

- $\mathbf{X}$ is the design matrix with first column of 1s and second column of x_i values
- $\beta = [\beta_0, \beta_1]^T$ is the parameter vector
- $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$ is the vector of observed outputs

For example, with 3 data points:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Step 7: Solve for the Parameters

Multiplying both sides by $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$:

$$\beta = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

For simple linear regression, this gives us:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(X, Y)}{Var(X)}$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

Where $\bar{x}$ and $\bar{y}$ are the means of x and y values respectively.

Geometric Interpretation

The least squares solution has an important geometric interpretation:

- The residuals are orthogonal (perpendicular) to the column space of $\mathbf{X}$
- The predicted values $\hat{\mathbf{y}}$ are the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$
- This is the closest point in the column space to the actual $\mathbf{y}$

Example: Calculating OLS Parameters

Consider this small dataset:

x	y
1	2
2	3
3	5

Let’s calculate: - $\bar{x} = \frac{1+2+3}{3} = 2$ - $\bar{y} = \frac{2+3+5}{3} = \frac{10}{3} \approx 3.33$ - $\sum(x_i - \bar{x})(y_i - \bar{y}) = (1-2)(2-3.33) + (2-2)(3-3.33) + (3-2)(5-3$
- $\sum(x_i - \bar{x})^2 = (1-2)^2 + (2-2)^2 + (3-2)^2 = 2$

Therefore: - $\beta_1 = \frac{2.67}{2} = 1.33$ - $\beta_0 = 3.33 - 1.33 \times 2 = 0.67$

Our fitted line is: $\hat{y} = 0.67 + 1.33x$

Summary: The Least Squares Method

Key points about the OLS derivation:

- We use calculus to find parameter values that minimize squared error
- The solution involves setting partial derivatives to zero
- The normal equations can be solved using matrix algebra
- The closed-form solution is $\beta = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}$
- For simple linear regression, parameters depend on means and covariances
- This approach generalizes to multiple regression with many predictors