Comprehensive Guide to Linear Regression Concepts with Jacobian and Hessian Matrices

Original Problem and Solution

Problem Statement

For a matrix A which has values as (1,1),(1,2),(1,3) and features as age and experience the target column value is salary which has values as \$2000, \$4000 and \$6000. If we take these features in matrix X, how to calculate $(X^TX)^{-1}X^T\mathbf{y}$?

Solution

Given:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 2000 \\ 4000 \\ 6000 \end{bmatrix}$$

Compute $\mathbf{a} = (X^TX)^{-1}X^T\mathbf{y}$:

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

2. $\det(X^TX) = 3 \cdot 14 - 6 \cdot 6 = 42 - 36 = 6$

$$(X^TX)^{-1} = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} = \begin{bmatrix} \frac{7}{3} & -1 \\ -1 & \frac{1}{2} \end{bmatrix}$$

3. $X^T\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2000 \\ 4000 \\ 6000 \end{bmatrix} = \begin{bmatrix} 12000 \\ 28000 \end{bmatrix}$

4. $\mathbf{a} = \begin{bmatrix} \frac{7}{3} & -1 \\ -1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 12000 \\ 28000 \end{bmatrix} = \begin{bmatrix} 28000 - 28000 \\ -12000 + 14000 \end{bmatrix} = \begin{bmatrix} 0 \\ 2000 \end{bmatrix}$

Thus, the linear regression model is:

$$\boxed{\text{Salary} = 0 + 2000 \cdot \text{Experience}}$$

Geometric Interpretation of Coefficients in X-Y Plane

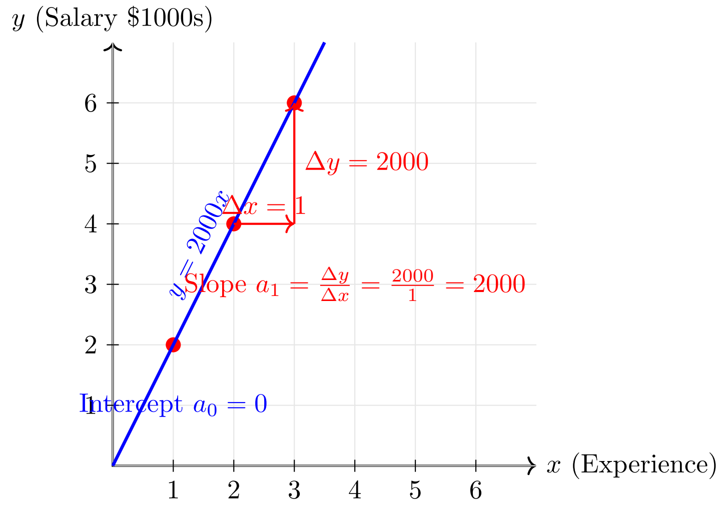
Representation of Coefficients and Slope

In the X-Y plane, the linear regression equation $y = a_0 + a_1x$ represents a straight line where:

- a_0 is the **y-intercept** (where the line crosses the y-axis)
- a_1 is the **slope** (steepness of the line)

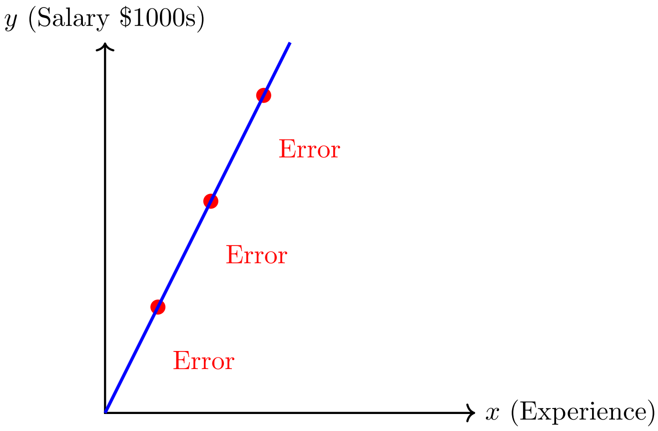
For our solution $\mathbf{a} = \begin{bmatrix} 0 \\ 2000 \end{bmatrix}$:

- **Intercept** $a_0 = 0$: The line passes through the origin (0,0)
- **Slope** $a_1 = 2000$: For each unit increase in Experience, Salary increases by \$2000



Geometric Meaning of the Solution

The coefficients $\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$ define the line that minimizes the sum of squared vertical distances between the data points and the line:



The optimal coefficients minimize the sum of these squared errors.

Connection to Jacobian and Hessian Matrices

Jacobian Matrix in Linear Regression

The Jacobian matrix represents the first derivatives of the cost function with respect to the parameters. For linear regression with mean squared error:

The cost function is:

$$J(\mathbf{a}) = \frac{1}{2m} \sum_{i=1}^{m} (h_{\mathbf{a}}(\mathbf{x}^{(i)}) - y^{(i)})^2 = \frac{1}{2m} \|\mathbf{X}\mathbf{a} - \mathbf{y}\|^2$$

The Jacobian (gradient) is:

$$\nabla J(\mathbf{a}) = \frac{1}{m} X^T (X\mathbf{a} - \mathbf{y})$$

For our solution $\mathbf{a} = \begin{bmatrix} 0 \\ 2000 \end{bmatrix}$:

$$\nabla J(\mathbf{a}) = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \left(\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 2000 \end{bmatrix} - \begin{bmatrix} 2000 \\ 4000 \\ 6000 \end{bmatrix} \right) = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The gradient is zero at the optimal solution, as expected.

Hessian Matrix in Linear Regression

The Hessian matrix represents the second derivatives of the cost function. For linear regression:

$$\nabla^2 J(\mathbf{a}) = \frac{1}{m} X^T X$$

For our problem:

$$\nabla^2 J(\mathbf{a}) = \frac{1}{3} \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & \frac{14}{3} \end{bmatrix}$$

The Hessian is positive definite (eigenvalues are positive), confirming that our solution is a minimum.

Clarification of Symbols and Function

Question

"So capital X is a feature matrix small a is a function and why is the output the value of this function is a vector am I correct?"

Answer

- $\mathbf{X}$: The **feature matrix** (input data). **Correct**.
- $\mathbf{a}$: The **parameter vector** (learned weights). This is *not* a function. **Incorrect**.
- $\mathbf{y}$: The **true target output vector**. **Correct**.
- $f(\mathbf{X})$: The **prediction function**. Its output is the vector of predictions, $\hat{\mathbf{y}}$. **Correct**.

The core equation of the linear model is:

$$\hat{\mathbf{y}} = f(\mathbf{X}) = \mathbf{X}\mathbf{a}$$

For our calculated values:

$$\hat{\mathbf{y}} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 2000 \end{bmatrix} = \begin{bmatrix} 2000 \\ 4000 \\ 6000 \end{bmatrix} = \mathbf{y}$$

The value of $\mathbf{a}$ is:

$$\boxed{\mathbf{a} = \begin{bmatrix} 0 \\ 2000 \end{bmatrix}}$$

The prediction function for a new data point is:

$$f(\mathbf{x}) = \mathbf{x} \cdot \mathbf{a} = a_0 + a_1 \cdot (\text{Experience}) = 2000 \cdot (\text{Experience})$$

Meaning of $X^T X$, $X^T y$, and $(X^T X)^{-1}$

Question

"What is the meaning of X transpose X and X transpose Y and X transpose X inverse, related with covariance correlation etc."

Answer

The Gram Matrix: $X^T X$

- **Size:** $(n \times n)$
- **Meaning:** Proportional to the **covariance matrix** of the features. The off-diagonals indicate feature correlation (multicollinearity).

$$\text{Covariance Matrix} \propto \frac{1}{m} X_c^T X_c$$

where X_c is the mean-centered feature matrix.

The Covariance Vector: $X^T y$

- **Size:** $(n \times 1)$
- **Meaning:** Proportional to the **covariance** between each feature and the target variable y.

$$\text{Covariance}(X_j, y) \propto (X^T y)_j$$

The Inverse Gram Matrix: $(X^T X)^{-1}$

- **Meaning:** The **precision matrix**. It adjusts for correlations between features, isolating their unique contributions.

The Complete Solution: $\mathbf{a} = (X^T X)^{-1} X^T y$

This formula calculates the optimal coefficients by:

1. Measuring the raw relationship between features and target $(X^T y)$.
2. Adjusting this relationship for the internal covariance structure of the features themselves $((X^T X)^{-1})$.

Geometric Interpretation in the (x, y) Plane

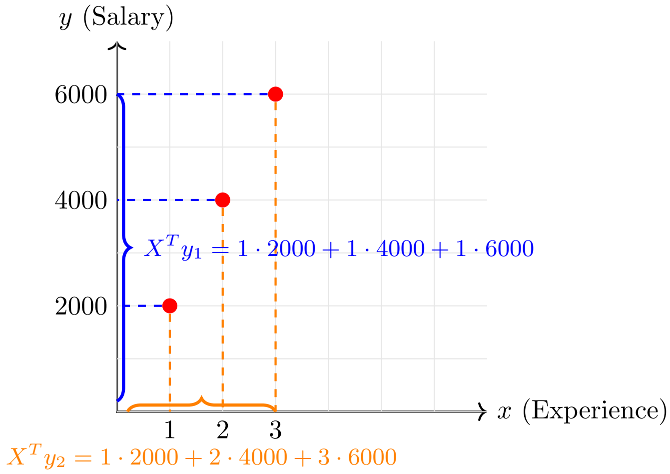
Question

"Please represent the meaning of X transpose X and X transpose Y and X transpose X inverse in x,y plane diagram or graph."

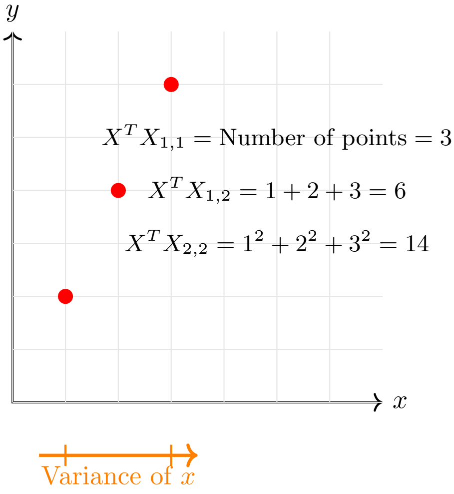
Answer

Using data points: (1, 2000), (2, 4000), (3, 6000)

1. Meaning of $X^T y$: Covariance with Target



2. Meaning of $X^T X$: Variance and Covariance of Features



3. Role of $(X^T X)^{-1}$: Scaling and Adjustment

The matrix $X^T X = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$ defines the "terrain" of the data. Its inverse:

$$(X^T X)^{-1} = \begin{bmatrix} \frac{7}{3} & -1 \\ -1 & \frac{1}{2} \end{bmatrix}$$

acts as a "gear ratio" that adjusts the raw signal $X^T y$ to find the optimal coefficients:

$$\mathbf{a} = (X^T X)^{-1} X^T y = \begin{bmatrix} \frac{7}{3} & -1 \\ -1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 12000 \\ 28000 \end{bmatrix} = \begin{bmatrix} 0 \\ 2000 \end{bmatrix}$$

4. Final Fitted Line