Linear Regression: Mathematical Notes

Let's consider a very small dataset for a simple linear regression problem, where we want to predict a continuous output variable based on a single input feature. This example will focus on **Ordinary Least Squares (OLS) Linear Regression** with a single feature.

Dataset: Advertising Spend vs. Sales

Suppose a small company tracked its weekly advertising spending (in thousands of dollars) and the corresponding weekly sales (in thousands of dollars). We want to understand the relationship and predict sales based on advertising spend.

Week	Advertising Spend (x, in \$1000)	Sales (y, in \$1000)
1	1	3
2	2	5
3	3	6
4	4	8

Goal: Find the best-fit linear model $y = \beta_0 + \beta_1 x$ that predicts Sales (y) based on Advertising Spend (x).

I. Ordinary Least Squares (OLS) Linear Regression

Linear Regression aims to model the relationship between a dependent variable (y) and one or more independent variables (x) by fitting a linear equation to observed data. In OLS, we find the line that minimizes the sum of the squared differences between the observed and predicted values (the residuals).

Model Equation:

For a simple linear regression with one independent variable:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where:

- y_i is the actual value of the dependent variable for the i-th observation.
- x_i is the value of the independent variable for the i-th observation.
- β_0 is the y-intercept (the predicted value of y when $x = 0$).
- β_1 is the slope (the change in y for a one-unit change in x).
- ϵ_i is the error term (or residual) for the i-th observation, representing the difference between the actual and predicted values.

Our goal is to estimate β_0 and β_1. The predicted value for y_i is denoted as $\hat{y}_i = \beta_0 + \beta_1 x_i$.

Objective Function (Least Squares):

We want to minimize the Sum of Squared Residuals (SSR) or Residual Sum of Squares (RSS):

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

To find the values of β_0 and β_1 that minimize RSS, we take partial derivatives with respect to β_0 and β_1, set them to zero, and solve the resulting system of equations.

Normal Equations (for simple linear regression):

The closed-form solutions for β_1 and β_0 are:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

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

where $\bar{x}$ is the mean of x values, and $\bar{y}$ is the mean of y values.

Calculations for Simple Linear Regression:

Step 1: Calculate necessary sums and means from the dataset. Number of data points (n) = 4

Week	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	3	1	9	3
2	2	5	4	25	10
3	3	6	9	36	18
4	4	8	16	64	32
Sum	$\sum x_i = 10$	$\sum y_i = 22$	$\sum x_i^2 = 30$	$\sum y_i^2 = 134$	$\sum x_i y_i = 63$

- $\sum x_i = 1 + 2 + 3 + 4 = 10$
- $\sum y_i = 3 + 5 + 6 + 8 = 22$
- $\sum x_i^2 = 1^2 + 2^2 + 3^2 + 4^2 = 1 + 4 + 9 + 16 = 30$
- $\sum y_i^2 = 3^2 + 5^2 + 6^2 + 8^2 = 9 + 25 + 36 + 64 = 134$
- $\sum x_i y_i = (1)(3) + (2)(5) + (3)(6) + (4)(8) = 3 + 10 + 18 + 32 = 63$

Calculate means:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{10}{4} = 2.5$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{22}{4} = 5.5$$

Step 2: Calculate β_1 (slope). Using the simplified formula:

$$\beta_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\beta_1 = \frac{(4)(63) - (10)(22)}{(4)(30) - (10)^2}$$

$$\beta_1 = \frac{252 - 220}{120 - 100}$$

$$\beta_1 = \frac{32}{20} = 1.6$$

Step 3: Calculate β_0 (y-intercept).

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

$$\beta_0 = 5.5 - (1.6)(2.5)$$

$$\beta_0 = 5.5 - 4.0 = 1.5$$

Step 4: Formulate the Regression Equation. The best-fit linear model is:

$$\hat{y} = 1.5 + 1.6x$$

Step 5: Make a prediction (e.g., if advertising spend is \$5000). If $x = 5$:

$$\hat{y} = 1.5 + 1.6(5) = 1.5 + 8.0 = 9.5$$

So, if the company spends \$5000 on advertising, the model predicts \$9500 in sales.

II. Multiple Linear Regression (Matrix Form)

For graduate-level machine learning, it's crucial to understand Linear Regression in its general **matrix form**, which extends seamlessly to multiple features.

Model Equation (Matrix Form):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where:

- $\mathbf{y}$ is an $n \times 1$ vector of dependent variable values ($[y_1, y_2, \ldots, y_n]^T$).
- $\mathbf{X}$ is an $n \times (p+1)$ design matrix. p is the number of features. The first column of $\mathbf{X}$ is typically a column of ones for the intercept term.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of regression coefficients ($[\beta_0, \beta_1, \ldots, \beta_p]^T$).
- $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of error terms ($[\epsilon_1, \epsilon_2, \ldots, \epsilon_n]^T$).

Objective Function (Matrix Form):

We minimize the squared Euclidean norm of the residuals:

$$\text{RSS} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Normal Equations (Matrix Form):

Differentiating RSS with respect to $\boldsymbol{\beta}$ and setting to zero yields the normal equations:

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

Solving for $\boldsymbol{\beta}$, we get the closed-form solution:

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

This solution exists if $\mathbf{X}^T\mathbf{X}$ is invertible (i.e., $\mathbf{X}$ has full column rank, which implies no perfect multicollinearity).

Calculations for Multiple Linear Regression (using our simple dataset as an example):

Even for our simple linear regression problem, we can demonstrate the matrix approach. Number of data points (n) = 4. Number of features (p) = 1.

Step 1: Define y and X matrices.

$$\mathbf{y} = \begin{pmatrix} 3 \\ 5 \\ 6 \\ 8 \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$$

Step 2: Calculate $\mathbf{X}^T\mathbf{X}$.

$$\mathbf{X}^T = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$$

$$= \begin{pmatrix} (1\cdot 1 + 1\cdot 1 + 1\cdot 1 + 1\cdot 1) & (1\cdot 1 + 1\cdot 2 + 1\cdot 3 + 1\cdot 4) \\ (1\cdot 1 + 2\cdot 1 + 3\cdot 1 + 4\cdot 1) & (1\cdot 1 + 2\cdot 2 + 3\cdot 3 + 4\cdot 4) \end{pmatrix}$$

$$= \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}$$

Step 3: Calculate $(\mathbf{X}^T\mathbf{X})^{-1}$. For a 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, the inverse is $\frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. Here, $a = 4, b = 10, c = 10, d = 30$. Determinant $ad - bc = (4)(30) - (10)(10) = 120 - 100 = 20$.

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{20}\begin{pmatrix} 30 & -10 \\ -10 & 4 \end{pmatrix} = \begin{pmatrix} 30/20 & -10/20 \\ -10/20 & 4/20 \end{pmatrix} = \begin{pmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{pmatrix}$$

Step 4: Calculate $\mathbf{X}^T\mathbf{y}$.

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix}\begin{pmatrix} 3 \\ 5 \\ 6 \\ 8 \end{pmatrix}$$

$$= \begin{pmatrix} (1\cdot 3 + 1\cdot 5 + 1\cdot 6 + 1\cdot 8) \\ (1\cdot 3 + 2\cdot 5 + 3\cdot 6 + 4\cdot 8) \end{pmatrix}$$

$$= \begin{pmatrix} 3 + 5 + 6 + 8 \\ 3 + 10 + 18 + 32 \end{pmatrix}$$

$$= \begin{pmatrix} 22 \\ 63 \end{pmatrix}$$

Step 5: Calculate $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{pmatrix}\begin{pmatrix} 22 \\ 63 \end{pmatrix}$$

$$= \begin{pmatrix} (1.5)(22) + (-0.5)(63) \\ (-0.5)(22) + (0.2)(63) \end{pmatrix}$$

$$= \begin{pmatrix} 33.0 - 31.5 \\ -11.0 + 12.6 \end{pmatrix}$$

$$= \begin{pmatrix} 1.5 \\ 1.6 \end{pmatrix}$$

So, $\beta_0 = 1.5$ and $\beta_1 = 1.6$. These are the same coefficients obtained from the simpler formulas, demonstrating the equivalence.

Discussion:

- **Assumptions of OLS:** Discuss the Gauss-Markov assumptions that guarantee OLS estimators are BLUE (Best Linear Unbiased Estimators): linearity, independence of errors, homoscedasticity, no perfect multicollinearity, and normally distributed errors (for inference).
- **Regularization (Ridge, Lasso, Elastic Net):** Explain how OLS can overfit when features are highly correlated or when $p > n$. Introduce regularization techniques (e.g., Ridge, Lasso) that add a penalty term to the RSS objective function, shrinking coefficients and improving generalization. This is a critical extension for graduate ML.
- **Computational Complexity:** Discuss the computational cost of $(\mathbf{X}^T\mathbf{X})^{-1}$ for large p. Inverting an $m \times m$ matrix takes approximately $O(m^3)$ time. For very large datasets or high-dimensional features, iterative methods like Gradient Descent or Stochastic Gradient Descent become necessary.
- **Gradient Descent:** Introduce Gradient Descent as an alternative optimization algorithm, particularly useful when the closed-form solution is computationally prohibitive or when the objective function is not convex (e.g., in non-linear models or neural networks).

- **Evaluation Metrics:** Beyond just finding coefficients, discuss how to evaluate a linear regression model: R^2, Adjusted R^2, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE).
- **Interpreting Coefficients:** Emphasize that in multiple linear regression, each β_i represents the change in y for a one-unit change in x_i, *holding all other features constant.* This interpretation is crucial.
- **Limitations:** Discuss when linear regression might not be appropriate (non-linear relationships, outliers, categorical outcomes) and lead into more advanced models.

Prepared By:
Md. Atikuzzaman
Lecturer
Department of Computer Science and Engineering
Green University of Bangladesh
Email: atik@cse.green.edu.bd