Machine Learning 101

Rajdeep Chatterjee, Ph.D.
Amygdala AI, Bhubaneswar, India [*]

January 2025

Linear regression

1 Introduction

Linear regression is a fundamental method in machine learning and statistics for modeling the relationship between a dependent variable and one or more independent variables. Ordinary Least Squares (OLS) is a common approach used to fit linear regression models by minimizing the sum of squared residuals.

2 Least Squares Method

The least squares method involves finding the parameters of a linear model such that the sum of the squared differences between the observed values and the predicted values is minimized. Mathematically:

$$J(\beta) = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{1}$$

where:

- y_i are the actual values.
- $\hat{y}_i = \beta_0 + \beta_1 x_i$ is the predicted value.
- m is the number of observations.

2.1 Objective

Minimize $J(\beta)$ to determine the optimal parameters β_0 and β_1.

[*] Amygdala AI, is an international volunteer-run research group that advocates for *AI for a better tomorrow* `http://amygdalaai.org/`.

3 Ordinary Least Squares (OLS)

OLS extends the least squares method for multiple variables. For a dataset X (design matrix) and target vector y, the model is:

$$\hat{y} = X\beta \tag{2}$$

The OLS solution minimizes:

$$J(\beta) = \|y - X\beta\|^2 \tag{3}$$

The closed-form solution is given by:

$$\beta = (X^T X)^{-1} X^T y \tag{4}$$

4 Toy Problem

Consider a dataset with two observations:

- Inputs $X = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$
- Outputs $y = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

Adding a bias term, the design matrix becomes:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \tag{5}$$

Using the OLS formula:

$$\beta = (X^T X)^{-1} X^T y \tag{6}$$

4.1 Solution

1. Compute $X^T X$: $\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^T \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$
2. Compute $X^T y$: $\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^T \begin{bmatrix} 2 \\ 3 \end{bmatrix}$
3. Compute β: $\begin{bmatrix} 3 & 5 \\ 5 & 13 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 8 \end{bmatrix}$

5 Challenges of OLS

- **Multicollinearity:** When predictors are highly correlated, $X^T X$ becomes near-singular.
- **Outliers:** OLS is sensitive to outliers, as they disproportionately influence the cost function.
- **Overfitting:** When the model has too many predictors, it may fit noise in the data rather than the underlying trend.

6 Gradient Descent Variants

6.1 Batch Gradient Descent

Updates β based on all training examples:

$$\beta := \beta - \alpha \nabla J(\beta) \tag{7}$$

6.2 Stochastic Gradient Descent (SGD)

Updates β based on a single example:

$$\beta := \beta - \alpha(y^{(i)} - \hat{y}^{(i)})x^{(i)} \tag{8}$$

6.3 Mini-batch Gradient Descent

Combines aspects of batch and stochastic gradient descent, updating β using a subset of examples.

7 Linear Regression with SGD: Step-by-Step

1. Initialize β to random values.
2. For each iteration:
   (a) Shuffle the dataset.
   (b) For each training example:
      - Compute prediction: $\hat{y} = x^T \beta$
      - Compute error: $e = \hat{y} - y$
      - Update parameters: $\beta := \beta - \alpha e x$

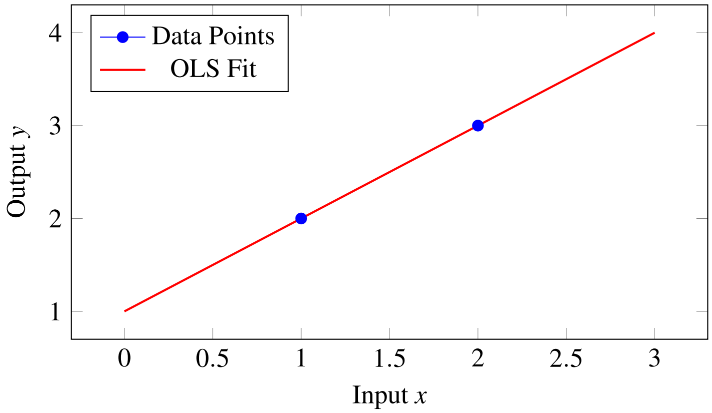

Figure 1: OLS fit for a toy problem.

8 Visualization

8.1 OLS Fit Example

9 Introduction to Linear Regression

Linear regression is a supervised learning algorithm used for modeling the relationship between a dependent variable y and one or more independent variables x. It aims to fit a linear equation to observed data.

The linear model is given by:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{9}$$

where:

- $\hat{y}$ is the predicted value of the dependent variable.
- $\theta_0, \theta_1, \ldots, \theta_n$ are the model parameters.
- $x_1, x_2, \ldots, x_n$ are the independent variables.

10 Cost Function for Linear Regression

The cost function quantifies the error between the predicted values ($\hat{y}$) and the actual values (y). For linear regression, we use the Mean Squared Error (MSE) as the cost function:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2 \tag{10}$$

where:

- m is the number of training examples.
- $\hat{y}^{(i)} = \theta^T x^{(i)}$ is the prediction for the i-th example.
- $y^{(i)}$ is the actual value for the i-th example.

The factor $\frac{1}{2}$ is included to simplify the derivative during gradient computation.

10.1 Deriving the Gradient of the Cost Function

To optimize $J(\theta)$, we compute its gradient with respect to θ_j:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) x_j^{(i)} \tag{11}$$

where $x_j^{(i)}$ is the j-th feature of the i-th example. This gradient guides us in updating θ to minimize the cost function.

11 Stochastic Gradient Descent (SGD)

Gradient Descent is an optimization algorithm that minimizes the cost function $J(\theta)$ by iteratively updating the model parameters.

11.1 Types of Gradient Descent

- **Batch Gradient Descent:** Uses the entire training dataset to compute the gradient at each iteration.
- **Stochastic Gradient Descent (SGD):** Uses a single training example to compute the gradient and update the parameters at each iteration.
- **Mini-Batch Gradient Descent:** Uses a small batch of training examples to compute the gradient at each iteration.

11.2 SGD Update Rule

For each training example $(x^{(i)}, y^{(i)})$, the parameters are updated as follows:

$$\theta_j := \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j} \tag{12}$$

Substituting the gradient, we get:

$$\theta_j := \theta_j - \alpha (\hat{y}^{(i)} - y^{(i)}) x_j^{(i)} \tag{13}$$

where α is the learning rate.

11.3 Advantages of SGD

- Faster updates as it uses only one example per iteration.
- Useful for large datasets.
- Can escape local minima due to its noisy updates.

11.4 Challenges of SGD

- Noisy updates can lead to fluctuations around the minimum.
- Requires careful tuning of the learning rate.
- Can take longer to converge compared to batch methods.

12 Toy Problem: Linear Regression with SGD

Let us solve a toy problem using SGD.

12.1 Problem Setup

Consider a dataset with one feature:

x	y
1	2
2	4
3	6
4	8

Table 1: Sample dataset.

12.2 SGD Steps

1. **Initialize Parameters:** Start with $\theta_0 = 0$ and $\theta_1 = 0$.
2. **Compute Predictions:** Use $\hat{y} = \theta_0 + \theta_1 x$.
3. **Update Parameters:** For each example $(x^{(i)}, y^{(i)})$:

$$\theta_0 := \theta_0 - \alpha(\hat{y}^{(i)} - y^{(i)}) \tag{14}$$

$$\theta_1 := \theta_1 - \alpha(\hat{y}^{(i)} - y^{(i)})x^{(i)} \tag{15}$$

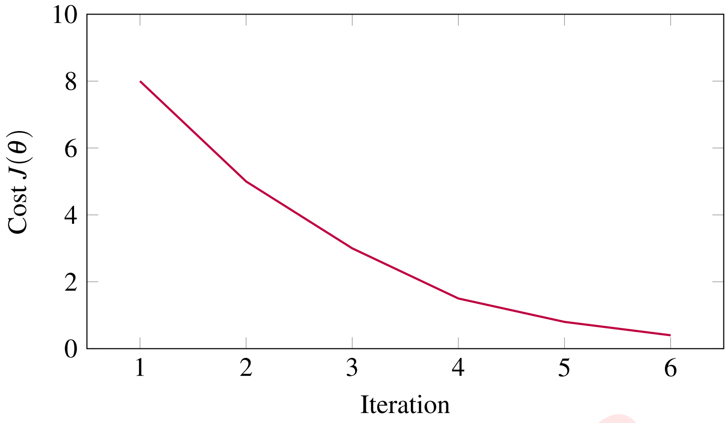

Figure 2: SGD convergence trajectory.

12.3 Visualization of SGD

13 Assumptions of Linear Regression

Linear regression relies on several assumptions to ensure the validity of its results. These assumptions, their verification methods, and remedies for violations are detailed below:

1. **Linearity:** The relationship between the independent variables (predictors) and the dependent variable (target) is linear.
   - **Verification:** Plot the observed data against the predicted values to check for a linear pattern.
   - **Remedies:** If the relationship is non-linear, consider applying transformations to the variables (e.g., log, square root) or using non-linear models.
2. **Independence:** Observations are independent, and residuals are not autocorrelated.
   - **Verification:** Use the Durbin-Watson test to detect autocorrelation in residuals.
   - **Remedies:** If autocorrelation exists, consider using time-series models or adding lagged variables.
3. **Homoscedasticity:** The variance of residuals is constant across all levels of the independent variables.
   - **Verification:** Create a residual vs. fitted values plot. Homoscedasticity is indicated by a random scatter with no discernible pattern.

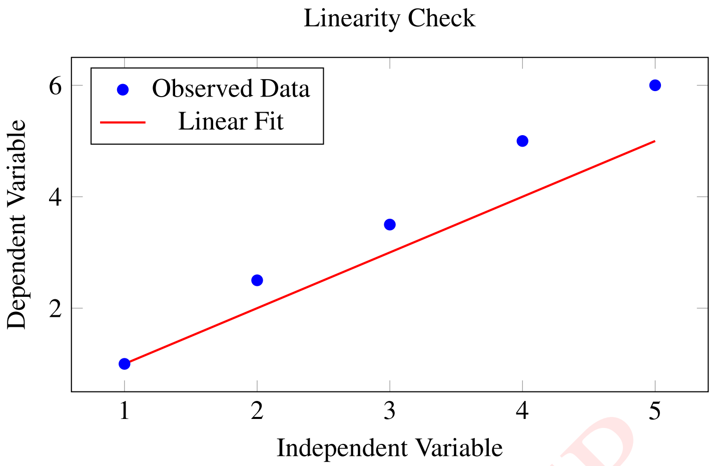

Figure 3: Checking the linear relationship between predictors and the target.

- **Remedies:** Use weighted least squares or transform the dependent variable if heteroscedasticity is present.

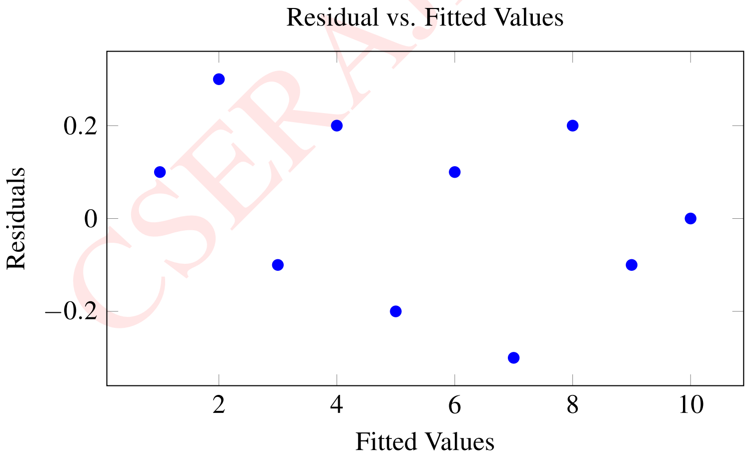

Figure 4: Residual plot to verify homoscedasticity. A random scatter indicates homoscedasticity.

4. **Normality of Residuals:** Residuals are normally distributed.
   - **Verification:** Use a Q-Q plot or the Shapiro-Wilk test to assess normality.
   - **Remedies:** Apply transformations to the target variable or use robust regression techniques if normality is violated.

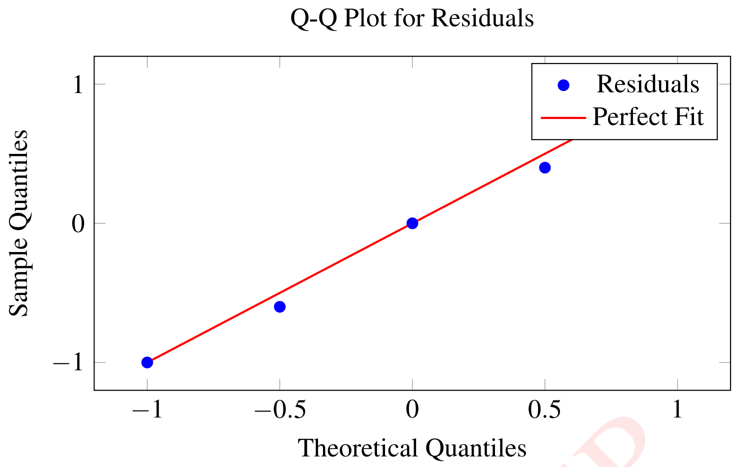

Figure 5: Q-Q plot to verify normality of residuals. Points close to the red line indicate normality.

5. **No Multicollinearity:** Independent variables are not highly correlated with each other.
   - **Verification:** Calculate the Variance Inflation Factor (VIF). A VIF greater than 10 indicates multicollinearity.
   - **Remedies:** Remove or combine highly correlated predictors, or use regularization techniques like ridge or lasso regression.
6. **Exogeneity:** Independent variables are uncorrelated with the error term.
   - **Verification:** Use the Hausman test to check for endogeneity.
   - **Remedies:** Include instrumental variables to address endogeneity.