LINEAR REGRESSION ALGORITHM

1. Definition:

Linear Regression is a fundamental statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The core idea is to establish a linear relationship between the input variables and the output. This method is widely used in predictive modeling and machine learning due to its simplicity and interpretability. By fitting a linear equation to observed data, Linear Regression helps in understanding how the dependent variable changes with respect to the independent variables.

2. Types of Data It Is Used On:

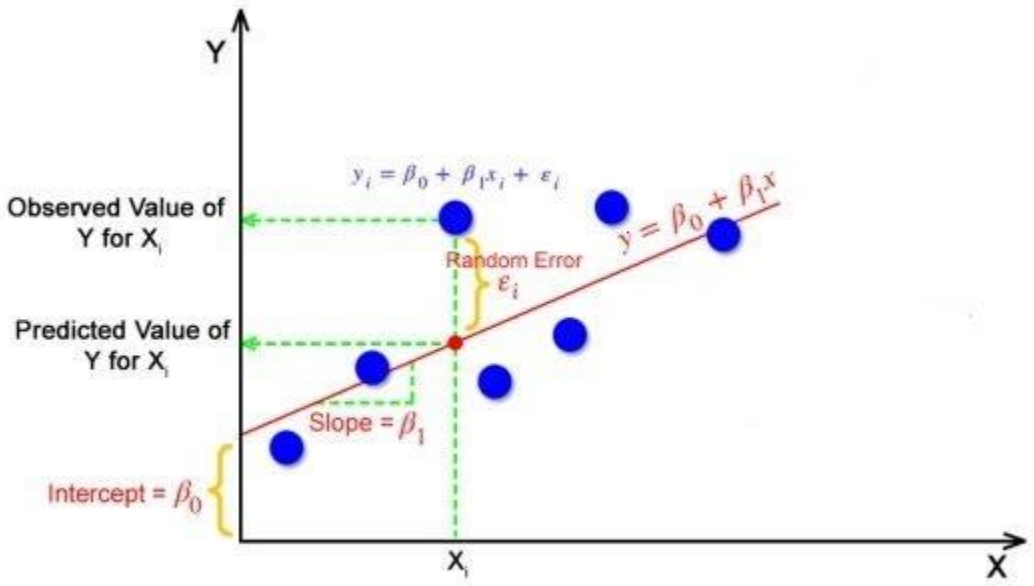
Linear Regression is primarily used on continuous target variables, such as house prices, stock prices, or any other measurable quantity. The independent variables can be either continuous or categorical (one-hot encoded). For instance, in predicting house prices, the size of the house (continuous) and the presence of a garage (categorical) can both be used as predictors.

3. Theory:

Linear Regression predicts the output variable (Y) by fitting a straight line (or hyperplane for multiple variables) to the data points. The equation of the line is represented as:

$$\boldsymbol{Y = \beta_0 + \beta_{1X_1} + \beta_{2X_2} + \ldots + \beta_{nX_n} + \epsilon}$$

Here,
$\boldsymbol{\beta_0}$ Is the intercept,
$\boldsymbol{\beta_1, \beta_2, \ldots \beta_n}$ are the coefficients (slopes of predictors),
$\boldsymbol{\epsilon}$ is the error term (residuals).



For example, in predicting house prices based on features like size and number of bedrooms, the equation might look like:

$$\boldsymbol{price} = \mathbf{20000} + \mathbf{3000}(\boldsymbol{size}) + \mathbf{1000}(\boldsymbol{bedrooms}) + \boldsymbol{\epsilon}$$

This equation indicates that for every additional square foot of size, the house price increases by $3000, and for each additional bedroom, the price increases by $1000, assuming all other factors remain constant.

4. Advantages:

Linear Regression offers several advantages:

- **Simplicity**: It is easy to understand and interpret, making it accessible to those without a deep statistical background.
- **Efficiency**: It is computationally efficient, allowing for quick analysis even with large datasets.
- **Interpretability**: The coefficients provide insights into the relationship between the predictors and the target variable.
- **Linearity**: It works well with linearly separable data, providing accurate predictions when the linearity assumption holds true.

5. Disadvantages:

Despite its advantages, Linear Regression has some limitations:

- **Sensitivity to Outliers**: Outliers can significantly distort the results, leading to inaccurate predictions.
- **Linearity Assumption**: It assumes a linear relationship between the predictors and the target, which may not always be the case in real-world data.
- **Multicollinearity**: When predictors are highly correlated with each other, it can lead to unstable estimates of the coefficients.
- **Overfitting**: Using too many predictors without regularization can lead to overfitting, where the model captures noise rather than the underlying pattern.

6. Mathematics and Statistics Behind It:

The mathematics behind Linear Regression involves minimizing the sum of squared errors (SSE):

$$\boldsymbol{SSE} = \sum_{\{\boldsymbol{i}=\mathbf{1}\}}^{\boldsymbol{n}} \left(\boldsymbol{y_i} - \left(\boldsymbol{\beta_0} + \boldsymbol{\beta_{1x_{\{1i\}}}} + \ldots + \boldsymbol{\beta_{nx_{\{ni\}}}} \right) \right)^2$$

Using Ordinary Least Squares (OLS), the best-fit line is found by calculating:

$$\boldsymbol{\beta} = \left(\boldsymbol{X^T X} \right)^{-1} \boldsymbol{X^T y}$$

Here,
(X) is the matrix of input variables,
(y) is the vector of target values.

This calculation ensures that the line of best fit minimizes the discrepancies between the observed and predicted values.

7. Real-Life Examples:

Linear Regression is used in various real-life scenarios. For example:

- Predicting house prices.
- Estimating sales revenue based on advertising budget.
- Forecasting temperature changes.
- Determining salary based on years of experience.
- Predicting patient health metrics like cholesterol levels.
- Stock market trend analysis.
- Demand forecasting in retail.
- Energy consumption prediction.
- Vehicle fuel efficiency estimation.
- Customer lifetime value prediction.

8. How to Use It in scikit-learn:

Using Linear Regression in scikit-learn is straightforward:

```python
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

# Sample data
X = [[1000], [1500], [2000], [2500], [3000]]  # House sizes
y = [200000, 250000, 300000, 350000, 400000]  # Prices

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and fit the model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Model coefficients
print("Intercept:", model.intercept_)
print("Coefficient:", model.coef_)
```

This example shows how to split data into training and testing sets, fit a Linear Regression model, and make predictions using scikit-learn.

9. How to Check Accuracy:

To evaluate the accuracy of a Linear Regression model, we can use metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 Score:

1. Mean Squared Error (MSE)

- **Definition:** Mean Squared Error (MSE) measures the average of the squares of the errors—that is, the average squared difference between the observed actual outcomes and the predicted outcomes.

- **Formula:**

$$MSE = \frac{1}{n}\sum_{\{i=1\}}^{n}\left(y_i - \{\widehat{y}\}_i\right)^2$$

where:
(n) is the number of data points.
$(y_{i)}$ is the actual value.
$\{\widehat{y}\}_i$ is the predicted value.

- **Interpretation:** MSE gives a higher weight to larger errors, making it sensitive to outliers. A *lower MSE indicates a better fit* of the model to the data. However, because it squares the errors, it can be difficult to interpret in the context of the original data.

- **Advantages:**
  - Penalizes larger errors more than smaller ones, making it useful for identifying models that make large errors.
  - Commonly used and well-understood metric.

- **Disadvantages:**
  - Sensitive to outliers, which can disproportionately affect the MSE.
  - Squaring the errors can make it difficult to interpret in the context of the original data.

2. Mean Absolute Error (MAE)

- **Definition:** Mean Absolute Error (MAE) measures the average of the absolute differences between the observed actual outcomes and the predicted outcomes.

- **Formula:**

$$MAE = \frac{1}{n}\sum_{\{i=1\}}^{n}\left|y_i - \{\widehat{y}\}_i\right|$$

where:
(n) is the number of data points.
$(y_{i)}$ is the actual value.
$\widehat{\{y\}_i}$ is the predicted value.

- **Interpretation:** MAE is less sensitive to outliers compared to MSE, as it does not square the errors. A *lower MAE indicates a better fit of the model* to the data. It is easier to interpret in the context of the original data because it represents the average error.

- **Advantages:**
  - Less sensitive to outliers compared to MSE.
  - Easier to interpret in the context of the original data.

- **Disadvantages:**
  - Does not penalize larger errors as much as MSE, which can be a drawback in some cases.

3. R^2 Score (Coefficient of Determination)

- **Definition:** The R^2 Score, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

- **Formula:**

$$R^2 = 1 - \frac{\left\{\sum_{\{i=1\}}^{n}\left(y_i - \{\widehat{y}\}_i\right)^2\right\}}{\left\{\sum_{\{i=1\}}^{n}\left(y_i - \overline{\{y\}}\right)^2\right\}}$$

where:
$(y_{i)}$ is the actual value
$\widehat{\{y\}_i}$ is the predicted value.
$\overline{\{y\}}$ is the mean of the actual values.

- **Interpretation:** An R^2 score of *1 indicates a perfect fit*, while an R^2 score of *0 indicates that the model does not explain any of the variance* in the dependent variable. *Negative values indicate that the model performs worse* than a horizontal line (mean of the actual values).

- **Advantages:**
  - Provides a measure of how well the independent variables explain the variance in the dependent variable.
  - Easy to interpret and widely used.

- **Disadvantages:**
  - Can be misleading if used alone, as it does not account for the number of predictors in the model.
  - Sensitive to outliers and can be inflated by overfitting.

```python
from sklearn.metrics import mean_absolute_error, r2_score

# Calculate metrics
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"Mean Absolute Error: {mae}")
print(f"R² Score: {r2}")
```

These metrics provide insights into how well the model performs, with lower MSE and MAE indicating better accuracy, and an R^2 score closer to 1 indicating a better fit.

10. Resources:

To further explore Linear Regression, here are some valuable resources:

towards-data-science/linear-regression-detailed-view
Stat Quest linear regression YouTube video
geeksforgeeks/mean-squared-error/
geeksforgeeks/how-to-calculate-mean-absolute-error-in-python/
freecodecamp/what-is-r-squared-r2-value-meaning-and-definition/
Linear regression YouTube video

11. Precautions:

When using Linear Regression, it's essential to take certain precautions. Ensure that the features are linearly related to the target variable, as non-linear relationships can lead to inaccurate predictions. Remove outliers, as they can distort the results and affect the model's performance. Normalize or standardize the data if the feature scales vary greatly, as this can improve the model's accuracy. Address multicollinearity by using techniques like Variance Inflation Factor (VIF) to identify and mitigate correlated predictors.