Regression: Simple Linear Regression, Logistic Regression, Mean Square Error

CSE3007 – Artificial Intelligence

GROUP 11

VIT Bhopal University
Bhopal–Indore Highway, Kothri Kalan,
Sehore
Madhya Pradesh – 466114

Presented to:
Dr. S. Ananthakumaran
Associate Professor

GROUP MEMBERS

RAHUL KUMAR	21BSA10068
ANURAG PRASAD	21BSA10075
SATYAM DEV YADUVANSHI	21BSA10076
SHREYANSH LAVANIA	21BSA10079
JAGRATI UPADHYAY	21BSA10082

CONTENTS

- WHAT IS REGRESSION?
- MOST COMMONLY USED REGRESSIONS.
- LINEAR REGRESSION
- ASSUMPTIONS OF LINEAR REGRESSION
- APPLICATIONS OF LINEAR REGRESSION
- ADVANTAGES AND DISADVANTAGES
- LOGISTIC REGRESSION
- ASSUMPTIONS OF LOGISTIC REGRESSION

CONTENTS

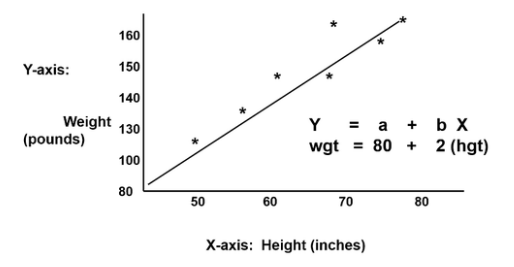
APPLICATIONS OF LOGISTIC REGRESSION
ADVANTAGES AND DISADVANTAGES
MEAN SQUARED ERROR

WHAT IS REGRESSION?

- Regression is a supervised machine learning technique which is used to predict continuous values.
- The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data.
- The three main metrics that are used for evaluating the trained regression model are variance, bias and error. If the variance is high, it leads to overfitting and when the bias is high, it leads to underfitting.
- Based on the number of input features and output labels, regression is classified as linear (one input and one output), multiple (many inputs and one output) and multivariate (many outputs).

WHAT IS REGRESSION?

- Imagine you have a bunch of balloons, each with a different weight written on it. You also have a measuring tape to measure their size. You blow them up and notice that bigger balloons tend to be heavier. Now, you want to create a game where you guess the weight of a balloon just by looking at its size.
- That's what regression does! It's like a smart assistant in this game. You show it some balloons with both weight and size written on them (this is your training data). It then tries to find a rule that connects size and weight. It might draw a line on a graph, like this:
- Now, when you show it a new balloon, it can use the line to guess its weight based on its size. The closer the balloon is to the line, the better the guess.
- This is a simple example of linear regression. There are other types of regression that can handle more complex relationships, like curves and multi-dimensional data. But the idea is always the same: using existing data to learn a rule that helps predict unknown values.



MOST COMMONLY USED REGRESSIONS.

- LINEAR REGRESSION
- LOGISTIC REGRESSION
- POLYNOMIAL REGRESSION
- STEPWISE REGRESSION
- RIDGE REGRESSION
- LASSO REGRESSION
- ELASTICNET REGRESSION

LINEAR REGRESSION

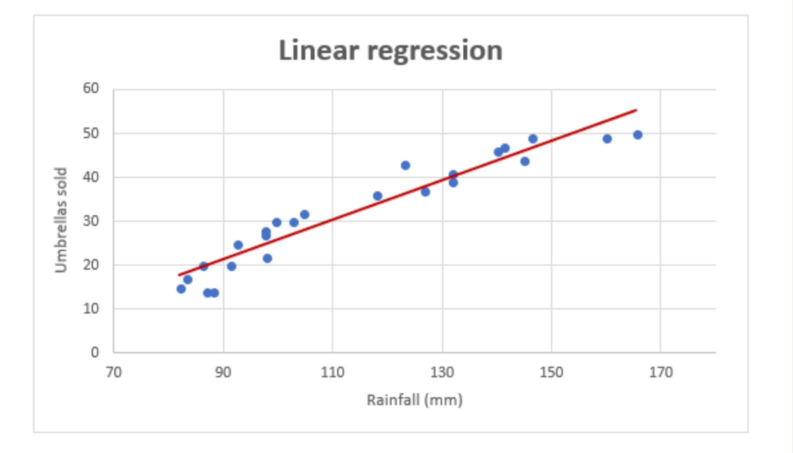
- Linear regression is a supervised machine learning technique used to predict a continuous value based on one or more input variables

- This is the simplest form of regression, and it assumes that there is a linear relationship between the dependent and independent variables. Linear regression is often used to predict continuous values, such as house prices, stock prices, and sales figures.

- The goal of linear regression is to find a linear relationship between the dependent variable (the output we want to predict) and the independent variable(s) (the input features).

LINEAR REGRESSION

As an example, let's take sales numbers for umbrellas for the last 24 months and find out the average monthly rainfall for the same period. Plot this information on a chart, and the regression line will demonstrate the relationship between the independent variable (rainfall) and dependent variable (umbrella sales):



LINEAR REGRESSION

Linear regression equation

Mathematically, a linear regression is defined by this equation:

$y = bx + a + \varepsilon$

- Where:
- *x* is an independent variable.
- *y* is a dependent variable.
- ***a*** is the ***Y–intercept***, which is the expected mean value of *y* when all *x* variables are equal to 0. On a regression graph, it's the point where the line crosses the Y axis.
- b is the ***slope*** of a regression line, which is the rate of change for *y* as *x* changes.
- ε is the random error term, which is the difference between the actual value of a dependent variable and its predicted value.

LINEAR REGRESSION

The linear regression equation always has an error term because, in real life, predictors are never perfectly precise. However, some programs, including Excel, do the error term calculation behind the scenes. So, in Excel, you do linear regression using the least squares method and seek coefficients a and b such that:

y = bx + a

For our example, the linear regression equation takes the following shape:

Umbrellas sold = b * rainfall + a

LINEAR REGRESSION

There exist a handful of different ways to find a and b. The three main methods to perform linear regression analysis in Excel are:

- Regression tool included with Analysis ToolPak
- Scatter chart with a trendline
- Linear regression formula

ASSUMPTIONS OF LINEAR REGRESSION

- Linear regression makes several assumptions about the data:

- Linearity: There is a linear relationship between the dependent and independent variables.
- Homoscedasticity: The variance of the residuals is constant across all values of the independent variable.
- Independence: The residuals are independent of each other.
- Normality: The residuals are normally distributed.
- If these assumptions are not met, the results of linear regression may be unreliable.

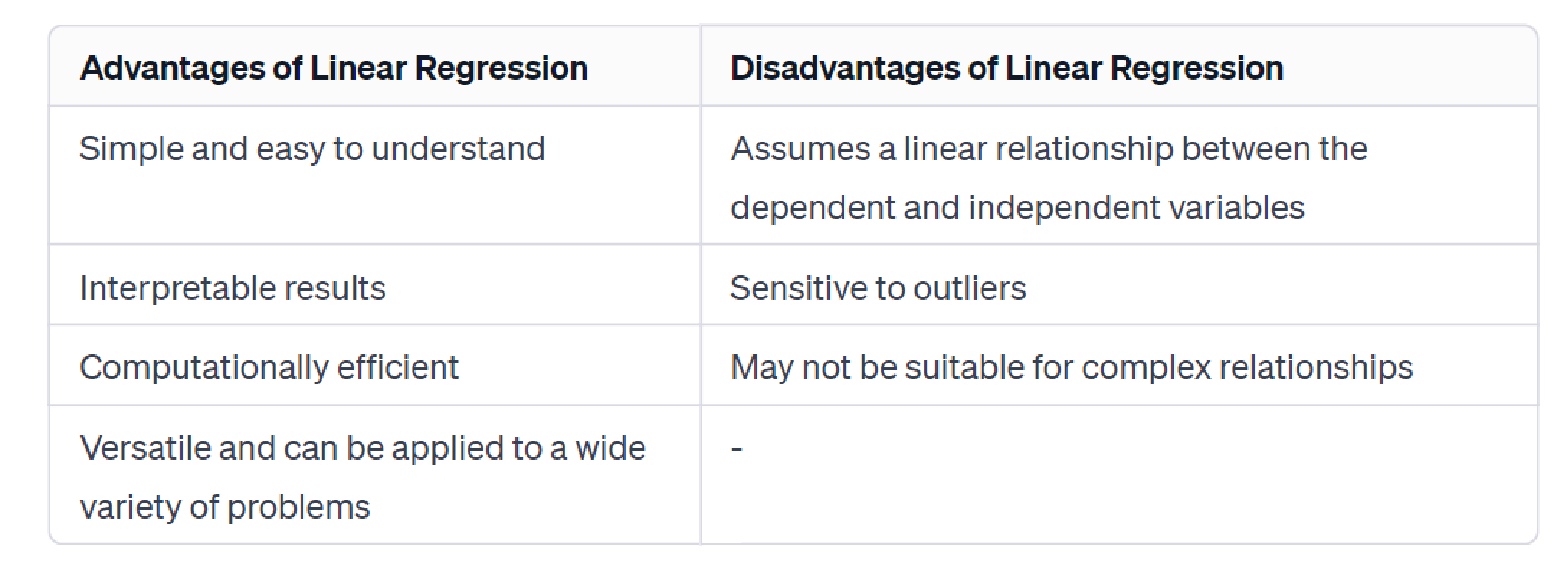
APPLICATIONS OF LINEAR REGRESSION:

- Linear regression is used for a wide variety of tasks, including:

- Predicting continuous values: Predicting house prices, stock prices, sales figures, etc.
- Correlation analysis: Identifying relationships between variables
- Trend analysis: Identifying trends in data over time
- Hypothesis testing: Testing hypotheses about relationships between variables

ADVANTAGES AND DISADVANTAGES

Let’s talk a look at the core advantages and disadvantages of Linear Regression:



Advantages of Linear Regression	Disadvantages of Linear Regression
Simple and easy to understand	Assumes a linear relationship between the dependent and independent variables
Interpretable results	Sensitive to outliers
Computationally efficient	May not be suitable for complex relationships
Versatile and can be applied to a wide variety of problems	-

LOGISTIC REGRESSION

- Logistic regression is a classification algorithm used to predict a categorical dependent variable (e.g., yes/no, success/failure) based on one or more independent variables. It is a statistical method that uses a mathematical equation to model the probability of an event occurring.

- Logistic regression transforms the dependent variable into a probability by applying a logistic function. The logistic function is a sigmoid function that maps any real number into the range [0, 1]. This means that the logistic function squashes any value into a probability between 0 and 1.

LOGISTIC REGRESSION

- The logistic function is defined as:
    - $\sigma(z) = 1 / (1 + \exp(-z))$
- where z is the dependent variable. The logistic function takes any real number as input and outputs a value between 0 and 1.

ASSUMPTIONS OF LOGISTIC REGRESSION

- Logistic regression makes several assumptions about the data:

- Multicollinearity: The independent variables should not be highly correlated with each other.
- Normality: The residuals should be normally distributed.
- Linearity: The relationship between the dependent variable and the independent variables should be linear.
- Independence: The residuals should be independent of each other.

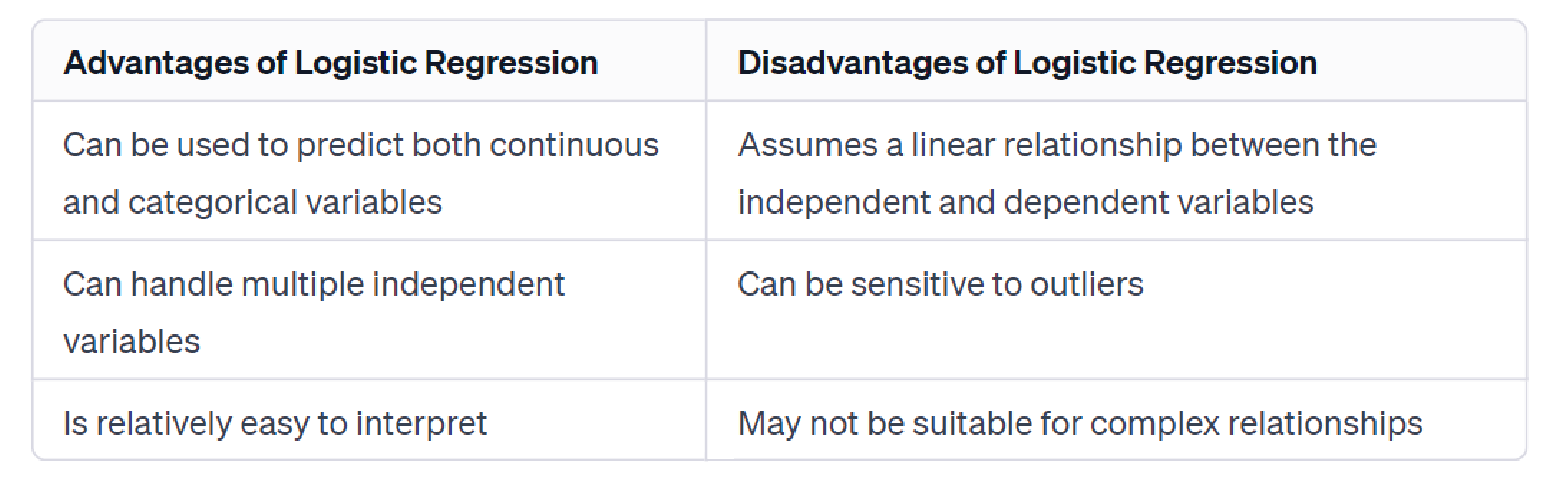
APPLICATIONS OF LOGISTIC REGRESSION:

- Logistic regression is used for a wide variety of tasks, including:

- Predicting customer churn: Predicting whether a customer is likely to churn (stop doing business) with a company.
- Fraud detection: Predicting whether a transaction is likely to be fraudulent.
- Medical diagnosis: Predicting whether a patient is likely to have a particular disease.
- Email spam filtering: Predicting whether an email is likely to be spam.

ADVANTAGES AND DISADVANTAGES

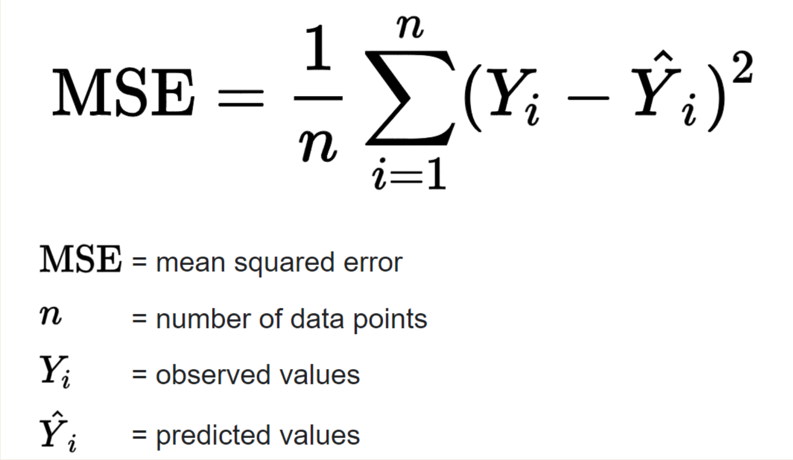
Let’s talk a look at the core advantages and disadvantages of Logistic Regression:



Advantages of Logistic Regression	Disadvantages of Logistic Regression
Can be used to predict both continuous and categorical variables	Assumes a linear relationship between the independent and dependent variables
Can handle multiple independent variables	Can be sensitive to outliers
Is relatively easy to interpret	May not be suitable for complex relationships

MEAN SQUARED ERROR

- Mean Squared Error (MSE) is a commonly used metric in the field of artificial intelligence, particularly in regression problems. It measures the average of the squared differences between predicted and actual values



$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

n = number of data points

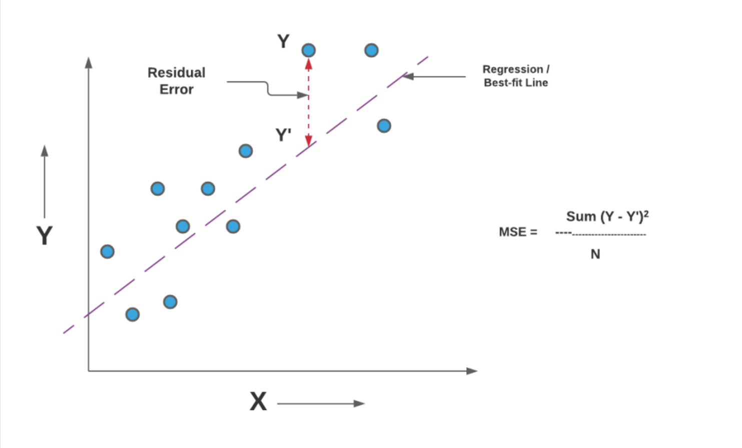
Y_i = observed values

$\hat{Y}_i$ = predicted values

MEAN SQUARED ERROR

- However, no model is perfect, and there will always be some error in the predictions. The mean squared error (MSE) is a way of measuring how much error there is in a model's predictions.

- The MSE is calculated by taking the average of the squared differences between the predicted values and the actual values



MEAN SQUARED ERROR

.

- Imagine you are trying to predict the temperature each day.
- You have the actual temperatures recorded, and you make predictions.

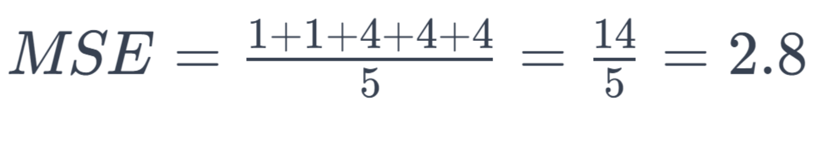
- Here are the actual temperatures (in Celsius) for five days:
- Actual Temperatures: 25, 28, 22, 30, 35

- Now, we make predictions:
- Predicted Temperatures: 24, 27, 20, 32, 33

- To calculate the MSE, we follow these steps:

MEAN SQUARED ERROR

- **Find the Differences:**
  - For each day, find the difference between the actual temperature and the predicted temperature.
- Differences: 1, 1, 2, –2, 2
- **Square the Differences:**
  - Square each of these differences to make sure they are all positive and to emphasize bigger errors.
- Squared Differences: 1, 1, 4, 4, 4
- **Find the Average:**
  - Add up these squared differences and find the average.



$$MSE = \frac{1+1+4+4+4}{5} = \frac{14}{5} = 2.8$$

MEAN SQUARED ERROR

.

- So, the Mean Squared Error (MSE) for our temperature predictions is 2.8.

- Interpretation: On average, the squared differences between the actual temperatures and the predicted temperatures are 2.8.

- The lower the MSE, the better the predictions match the actual values.

Thanks!

Do you have any questions?