Linear Regression

# Overview and Application

## Model Introduction

Linear Regression is a statistical method, foundational to machine learning, that is used to analyze the relationship between a dependent variable and one or more independent variables.

The goal is to find the **line of best fit**, a straight line that predicts output values within a range.

## Use Cases

The method is used in almost all industries and areas of academia such as economics, business and finance to forecast sales, prices of assets, assess risk or in medicine to predict the progression of disease.

## When and Why to use

Linear regressions is best used for:

**Nature of Relationship**

* **Linear relationships –** The data (via a scatter plot) shows the presence of a linear relationship.

**Type of Data**

* **Continuous Data –** The data is continuous and not categorical (although categorical data can be used if appropriately encoded, see one-hot encoding)
* **Time Series with Linear Trend –** If a time series shows a linear trend.

**Problem Type**

* **Prediction of Quantitative Values –** If the goal is to predict a continuous value (sales, temperature etc)
* **Understanding Relationships –** If the objective is to understand the strength and direction of the relationship between two variables.

**Data Size and Complexity**

* **Moderate-sized Datasets –** Linear regression *can* handle large datasets, but for extremely large datasets, more complex models or sampling techniques are usually more efficient. Smaller datasets tend to not to show trends so neatly.
* **Low to Moderate Dimensionality –** Linear regression can handle multiple predictors but runs into issues such as multicollinearity

# Mathematical Foundations

## Theory

Linear regression is based on the straight line equation:

Where y is the dependent variable, x the independent variable, m is the coefficient between them ( or slope) and b is the y-intercept.

## Algorithms

The most common method to find the best fit line is the least squares method which minimizes the sum of the squared differences (errors) between the observed and predicted values.

## Probabilistic Framework

Linear regression can be viewed from a probabilistic perspective, assuming that the regression errors are normally distributed.

# Model Architecture and Components

## Model Structure

In simple linear regression, there is one independent variable. The coefficient represents the change in the dependent variable for a one-unit change in the independent variable.

## Parameters and Hyperparameters

The parameters are the coefficients and intercept determined during model training. Hyperparameters include the choice of regularization and the solver used.

# Assumptions and Limitations

## Assumptions of the Data and Problem

Linearity, Independence, Homoscedasticity, No multicollinearity, and Normality of residuals are key assumptions.

## Understanding Limitations and Pitfalls

Limitations include sensitivity to outliers and potential for overfitting with many variables

# Model Building and Training

## Data Preparation

Data may need to be normalized.

Categorical data should be encoded, normally via one-hot encoding.

## Training Techniques

Training should be conducted on a training dataset then validate on a separate dataset.

# Model Evaluation

## Evaluation Metrics

The normal evaluation metrics of Linear Regression are:

* R-Squared
* Adjusted R-Squared
* Mean Squared Error
* Root Mean Squared Error

## Validation Techniques

The train-test split is the most commonly used validation method.

# Model Diagnostics and Refinement

## Overfitting vs. Underfitting

Regularization techniques like ridge and lasso help prevent overfitting.

## Interpretability

Coefficients in the model directly represent the impact of predictors.

# Advanced Topics and Extensions

## Ensemble Methods

Techniques like boosting can be applied to linear regression.

## Transfer Learning

na

## Regularization

Ridge, Lasso and Elasticnet are extensions of the model.

# Ethical Considerations

## Bias and Fairness

## Transparency and Accountability

# Continuous Learning

## Research

## Community