# Module 2

# Regression Analysis

## 📌 Introduction to Regression

Regression is a fundamental supervised learning technique used to model the relationship between independent variables (features) and a continuous dependent variable (target). This technique is crucial for predictive modeling, forecasting, and trend analysis across industries such as finance, healthcare, sales, and engineering.

### 🔹 General Regression Equation

Regression models follow a general mathematical equation:

Where:

* → Predicted value (target variable)
* X → Independent variables (predictors/features)
* → Error term (unexplained variance)

## 🔹 Why is Regression Important?

✔ Helps in understanding relationships between features and outcomes.  
✔ Facilitates interpretability and explainability of models.  
✔ Used for decision-making, forecasting, and optimization.  
✔ Forms the foundation for advanced ML techniques like GLMs and deep learning regressors.

## 🔹 Correlation & Feature Engineering in Regression

### 1️. Feature-Target Correlation

* Each independent variable should be strongly correlated with the target variable.
* A higher correlation means the feature has predictive power.
* **Pearson’s Correlation Coefficient (r)** helps quantify this relationship:
  + r > 0.7→ Strong correlation.
  + 0.3 ≤ r ≤ 0.7→ Moderate correlation.
  + r < 0.3→ Weak or no correlation.

### 2. Feature-to-Feature Correlation (Multicollinearity)

* Features should **not** be highly correlated with each other.
* **Multicollinearity** can cause unstable coefficient estimates, making model interpretation difficult.

#### Detection:

* Correlation matrix (remove features with |r| > 0.8).
* Variance Inflation Factor (VIF) (remove features with VIF > 10).

#### Solutions:

#### ✔ Remove redundant variables. ✔ Use **Principal Component Analysis (PCA)** or **Regularization (Lasso/Ridge)**.

## 🔹 Types of Regression Models & Their Applications

### Simple Linear Regression (SLR)

* Models the relationship between **one predictor** and the target variable.
* Formula where:
  + → Intercept (baseline prediction)
  + → Slope (rate of change)
  + **X** → Predictor variable

✔ Best for clear **linear** relationships.  
✔ **Fails** when relationships are non-linear.  
✔ Assumes **homoscedasticity** (constant variance in residuals).

🟢 **Example:** Predicting **CO₂ emissions** based on engine size.

### Multiple Linear Regression (MLR)

* Extends SLR by incorporating **multiple predictors**.
* Formula:
* Useful for **non-linear** trends.  
  ✔ Risk of **overfitting** with high-degree polynomials.  
  ✔ **Cross-validation and regularization** help prevent overfitting.
* 🟢 **Example:** Modeling **fuel efficiency trends**.

### Polynomial Regression (MLR)

* Introduces **higher-order** terms to capture curvature.
* Formula:

✔ Useful for **non-linear** trends.  
✔ Risk of **overfitting** with high-degree polynomials.  
✔ **Cross-validation and regularization** help prevent overfitting.

🟢 **Example:** Modeling **fuel efficiency trends**.

### Non-Linear Regression

* Used when relationships cannot be captured by a straight line or polynomial function.
* **Exponential Regression**

✔ Used for **population growth, investment returns**.

* **Logarithmic Regression**

✔ Used for **diminishing returns, resource consumption**

* **Sinusoidal Regression**

✔ Used for **seasonal trends, energy demand forecasting**.

## 🔹 Evaluating Regression Models

To assess model performance, use the following metrics

### 1️. Mean Squared Error (MSE)

* Measures average squared error between actual and predicted values:

✔ Lower **MSE** = **Better model fit**.

### 2. Root Mean Squared Error (RMSE)

* Square root of MSE, easier to interpret:

### 3. Coeficient of Determination ( Score)

* Measures how well the model explains variance in data:

✔  **closer to 1** = **strong model**.

✔  **near 0** = poor model.