

Assignment Number 02

Name: Mihir Unmesh Patil

Roll No: TYCOC213

Batch: C/C-3

Aim

The primary objective of this assignment is to analyze the relationship between input features and output variables using **regression models**. The study involves:

1. **Linear and Polynomial Regression** to analyze the relationship between **hours spent driving** and **risk score**.
2. **Comparison of regression models** (Linear Regression, Ridge, Lasso, and ElasticNet) on the **Diabetes dataset**.
3. **Performance evaluation** of models using **Mean Squared Error (MSE)** and **R² Score**.
4. **Visualization of model predictions, residuals, and feature importance** for better interpretability.

Objectives

- Implement **Linear and Polynomial Regression** for predicting risk scores based on driving hours.
- Apply **StandardScaler** for feature normalization in the diabetes dataset.
- Train and compare multiple regression models (Linear, Ridge, Lasso, and ElasticNet).
- Compute **performance metrics (MSE, R² Score)** for evaluating regression models.
- Analyze model residuals and feature importance for interpretation.
- Visualize results using plots for better insights.

Theory

Regression Analysis

Regression is a statistical technique used to model relationships between a dependent variable (y) and one or more independent variables (X). It is widely used in **predictive modeling** to estimate unknown values based on known data.

Linear Regression

Linear Regression finds the best-fitting straight line through data points using the equation:

$$y = mX + c$$

where:

- y is the predicted value,
- m is the slope (coefficient),
- X is the independent variable,
- c is the intercept.

The model minimizes the **Mean Squared Error (MSE)** to find the optimal m and c .

Polynomial Regression

Polynomial Regression extends Linear Regression by introducing polynomial terms:

$$y = a_0 + a_1X + a_2X^2 + a_3X^3 + \dots + a_nX^n$$

where higher-degree terms allow for modeling **non-linear** relationships in data.

Regularized Regression Models

Regularization techniques are used to **prevent overfitting** by adding penalty terms to the loss function. The following methods were implemented:

1. **Ridge Regression (L2 Regularization)**
 - Adds a **penalty on large coefficients** using **L2 norm**:
2. $Loss = \sum (y - \hat{y})^2 + \lambda \sum \theta^2$
 - Helps in reducing overfitting but does not eliminate features entirely.
3. **Lasso Regression (L1 Regularization)**
 - Uses **L1 norm** for penalty:
4. $Loss = \sum (y - \hat{y})^2 + \lambda \sum |\theta|$
 - Shrinks some coefficients to **zero**, performing **feature selection**.
5. **ElasticNet Regression**
 - Combines L1 and L2 penalties:
6. $Loss = \sum (y - \hat{y})^2 + \lambda_1 \sum |\theta| + \lambda_2 \sum \theta^2$
 - Offers the benefits of both Ridge and Lasso.

Performance Metrics

The models were evaluated using:

1. Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2$$

- Measures how close predictions are to actual values. Lower MSE indicates better fit.

2. R² Score (Coefficient of Determination)

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}}$$

$$R^2 = 1 - \frac{SS_{\text{total}} - SS_{\text{residual}}}{SS_{\text{total}}}$$

- Indicates how well independent variables explain the variance in the dependent variable. Higher R² is better.

Results

CODE:

 Assignment_02_ML.ipynb

Part 1: Driving Hours vs Risk Score Analysis

We applied **Linear Regression** and **Polynomial Regression (degree=3)** to analyze the relationship between **hours spent driving** and **risk score**.

Model	R ² Score
Linear Regression	0.437
Polynomial Regression	0.808

- Linear Regression** resulted in an **R² score of 0.437**, indicating a moderate correlation.
 - Polynomial Regression (degree=3)** achieved a much higher **R² score of 0.808**, suggesting that a nonlinear model fits the data better.
 - Visualization** showed that the polynomial curve captured patterns in the data more effectively than a straight-line fit.
-

Part 2: Diabetes Dataset - Model Comparison

We compared **Linear Regression**, **Ridge**, **Lasso**, and **ElasticNet** on the **Diabetes dataset** after standardizing the features.

Model	MSE	R ² Score
Linear Regression	2900.19	0.453
Ridge Regression	2892.01	0.454
Lasso Regression	2824.56	0.467
ElasticNet	2888.70	0.455

Observations:

- **Lasso Regression had the lowest MSE (2824.56)** and the highest R² Score (0.467), making it the best-performing model.
- **Ridge Regression** and **ElasticNet** performed slightly better than **Linear Regression**, but not as well as Lasso.
- **Feature importance analysis** showed that Lasso regression effectively eliminated less relevant features, improving model efficiency.

Residual Analysis

Residual plots were created to check for any patterns in prediction errors:

- The **residuals were randomly distributed**, indicating that the models performed reasonably well.
- No significant heteroscedasticity (patterned residuals) was observed, which confirms the validity of the regression assumptions.

Feature Importance Analysis

- Feature importance was plotted for **Linear, Ridge, Lasso, and ElasticNet models**.
 - **Lasso Regression eliminated some features**, suggesting that certain attributes in the diabetes dataset had low predictive value.
-

Conclusion

This assignment demonstrated the effectiveness of various **regression techniques** for **predictive modeling**.

Key Takeaways

1. **Driving Hours vs Risk Score Analysis**
 - **Polynomial Regression (degree=3)** outperformed **Linear Regression**, proving that a non-linear approach is better suited for modeling **driving risk based on hours driven**.
2. **Diabetes Dataset Model Comparison**
 - **Lasso Regression** provided the best performance with **lowest MSE and highest R^2 score**, suggesting that some features were unnecessary.
 - **Ridge Regression and ElasticNet** performed slightly better than **Linear Regression**, showing that regularization helps improve model performance.
 - **Feature selection played a crucial role** in improving the predictive power of models.
3. **Practical Implications**
 - **Polynomial Regression** can be useful for modeling relationships that are not purely linear, such as risk assessments in **automotive safety**.
 - **Lasso Regression** is beneficial when working with **high-dimensional datasets**, such as medical records, where selecting the most relevant features is important.
 - **Regularization techniques (Lasso, Ridge, ElasticNet)** prevent overfitting, making models more generalizable to unseen data.