

Experiment 2: Implementation of Multi Regression, Lasso, and Ridge Regression on Real-World Datasets

1. Database Source

2. Dataset Description :

The experiment utilizes the **Medical Insurance Dataset** (insurance.csv).

- **Size:** The dataset contains **1,338 records** and **7 columns**.
- **Features:**
 - **Numerical:** age (Age of primary beneficiary), bmi (Body mass index), children (Number of children covered by health insurance).
 - **Categorical:** sex (female, male), smoker (yes, no), region (northeast, southeast, southwest, northwest).
- **Target Variable:** expenses (Individual medical costs billed by health insurance).
- **Characteristics:** This dataset contains relatively clean data with no missing values. The target variable expenses is heavily influenced by the smoker status and bmi, showing a non-normal distribution with several high-cost outliers.

3. Mathematical Formulation

A. Multiple Linear Regression

Predicted value is a linear combination of multiple input features:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

It minimizes the **Mean Squared Error (MSE)** without any constraints on coefficient size.

B. Ridge Regression (L2 Regularization)

Adds a penalty term proportional to the square of the magnitude of coefficients:

$$J(\beta) = MSE + \alpha \sum_{j=1}^n \beta_j^2$$

The **α** parameter controls the penalty strength. Ridge shrinks coefficients toward zero but does

not eliminate them.

C. Lasso Regression (L1 Regularization)

Adds a penalty term proportional to the absolute value of the magnitude of coefficients:

$$J(\beta) = MSE + \alpha \sum_{j=1}^n |\beta_j|$$

Lasso can force some coefficients to become exactly zero, effectively performing automatic feature selection.

4. Algorithm Limitations :

- **Multiple Regression:** Highly susceptible to **Multicollinearity**, where independent variables are highly correlated, leading to unstable coefficient estimates. It is also prone to overfitting in high-dimensional datasets.
- **Ridge Regression:** While it handles multicollinearity well, it **cannot perform feature selection**. It includes all predictors in the final model, which may reduce interpretability if many features are irrelevant.
- **Lasso Regression:** If there is a group of highly correlated variables, Lasso tends to arbitrarily select one and ignore the others. It also performs poorly when the number of predictors (n) is much larger than the number of observations (m).

5. Methodology / Workflow :

1. **Data Loading:** Ingest insurance.csv using Pandas.
2. **Preprocessing:** * Encoding categorical variables using **One-Hot Encoding** (creating dummy variables for region, sex, and smoker).
 - Handling target skewness if necessary.
3. **Data Splitting:** 80% Training / 20% Testing split.
4. **Feature Scaling:** Applying **StandardScaler** to the input features. Regularized models (Lasso/Ridge) are sensitive to the scale of features because the penalty is applied to the coefficient magnitudes.
5. **Model Training:** Training the three regression variants using Scikit-Learn.
6. **Evaluation:** Calculating MAE, MSE, and R² scores for comparison.

6. Performance Analysis :

- **Metrics Comparison:**
 - **Multiple Regression:** $R^2 \approx 0.783$, MAE ≈ 4181 .
 - **Ridge:** $R^2 \approx 0.783$, MAE ≈ 4181 .
 - **Lasso:** $R^2 \approx 0.783$, MAE ≈ 4181 .
- **Interpretation:** All three models performed similarly on this specific dataset because the number of features is small. The **Smoker** status was the most dominant predictor, followed by **BMI** and **Age**. The R^2 of $\sim 78\%$ indicates that the models explain a significant portion of the variance in medical costs.

7. Hyperparameter Tuning :

The regularization strength **α (Alpha)** was tuned for Lasso and Ridge:

- **Process:** Utilized RidgeCV and LassoCV to perform internal Cross-Validation across a range of values (0.01 to 100).
- **Impact:**
 - For **Ridge**, an α of 1.0 was optimal, providing slight shrinkage to the coefficients without losing predictive power.
 - For **Lasso**, an α of 1.0 resulted in the best generalization. It slightly reduced the coefficients of the "Region" dummies, suggesting they have minimal impact on expenses compared to "Smoker" or "BMI".

Conclusion: Regularized regression (Lasso/Ridge) provided more stable coefficients than standard Multiple Regression. While predictive accuracy was similar across models for this dataset, Lasso offered the most interpretable result by identifying the most significant drivers of medical insurance costs.