Linear Regression Analysis

CS4372 - Applied Machine Learning

Authors: Roman Hauksson-Neill, Ivan Masyuk

Dataset: Combined Cycle Power Plant Data Set (UCI ML Repository)

Abstract

This report presents a comprehensive linear regression analysis of a Combined Cycle Power Plant dataset from the UCI Machine Learning Repository. We implemented two regression models: Stochastic Gradient Descent (SGD) using scikit-learn and Ordinary Least Squares (OLS) using statsmodels. The analysis includes extensive data preprocessing, feature selection based on statistical significance and multicollinearity analysis, and model comparison. Our results demonstrate strong predictive performance with R^2 values exceeding 0.92 for both approaches.

1. Dataset Description

1.1 Data Source and Context

The Combined Cycle Power Plant dataset contains 9,568 hourly average ambient variables recorded at a combined cycle power plant over 6 years (2006-2011). The dataset is publicly available through the UCI Machine Learning Repository at: https://archive.ics.uci.edu/static/public/294/data.csv

1.2 Variable Description

The dataset contains five numerical variables:

- **Ambient Temperature (°C):** Temperature of the surrounding environment
- **Vacuum (cm Hg):** Exhaust vacuum measurement
- **Ambient Pressure (millibar):** Atmospheric pressure
- **Relative Humidity (%):** Moisture content in the air
- **Power Output (MW):** Net hourly electrical energy output (target variable)

1.3 Data Characteristics

After removing duplicate feature combinations, the final dataset contains 9,527 observations. All variables are continuous numerical values with no missing data. The power output ranges from 420.26 MW to 495.76 MW with a mean of 454.34 MW.

2. Data Preprocessing and Analysis

2.1 Data Consistency and Quality

Initial analysis revealed 82 rows with duplicate feature combinations that produced identical outputs, which were safely removed to eliminate redundancy. No missing values were detected, and all variables maintained consistent numerical data types throughout the analysis.

2.2 Distribution Analysis

Shapiro-Wilk normality tests ($\alpha = 0.05$) revealed that all variables significantly deviate from normal distribution ($p < 0.001$). This non-normality is expected in industrial datasets due to operational constraints and control systems that maintain variables within specific operational ranges.

2.3 Standardization Rationale

Standardization was essential due to dramatically different scales:

- Ambient pressure: 1013 millibar (large magnitude)
- Vacuum: 54 cm Hg (medium magnitude)
- Temperature: 20°C (small magnitude)

- Humidity: 73% (medium magnitude)

Without standardization, ambient pressure would dominate SGD optimization due to its large numerical scale. StandardScaler was applied to achieve zero mean and unit variance for all features.

2.4 Correlation Analysis

Comprehensive correlation analysis revealed strong relationships between features and the target variable:

Feature-Target Correlations:

- Ambient Temperature: −0.948 (strong negative correlation)
- Vacuum: −0.870 (strong negative correlation)
- Ambient Pressure: +0.519 (moderate positive correlation)
- Relative Humidity: +0.391 (moderate positive correlation)

Multicollinearity Concerns: The strongest inter-feature correlation is between ambient temperature and vacuum (r = 0.844), indicating potential multicollinearity. This high correlation suggests these variables measure related aspects of the power plant's operating environment.

2.5 Feature Selection Strategy

We systematically evaluated all possible feature combinations (15 total) using linear regression to identify optimal subsets. The analysis considered both predictive performance (R^2) and statistical reliability (VIF, condition number, p-values).

Performance Results:

Features	**Test R^2**	**Trustworthiness**
All 4 features	0.9284	✗ (VIF=5.89)
Temp + Vacuum + Humidity	0.9282	✓ (VIF=4.88)
Temp + Pressure + Humidity	0.9209	✓ (VIF=2.01)
Temp + Humidity	0.9209	✓ (VIF=1.41)
Temp + Vacuum + Pressure	0.9179	✓ (VIF=3.81)

Key Findings:

- Using all features achieves highest R^2 (0.9284) but violates VIF threshold (>5)
- Best trustworthy combination: Temperature + Vacuum + Humidity (R^2 = 0.9282, VIF = 4.88)
- Most stable combination: Temperature + Pressure + Humidity (R^2 = 0.9209, VIF = 2.01)
- Ambient temperature is the dominant predictor (single-feature R^2 = 0.9000)

3. Model Implementation Strategy

3.1 SGD Regression Approach

The SGD implementation uses scikit-learn's SGDRegressor with comprehensive hyperparameter tuning:

Hyperparameter Grid:

- Learning rate: ['constant', 'optimal', 'invscaling', 'adaptive']
- Penalty: ['l2', 'l1', 'elasticnet']
- Alpha: [0.0001, 0.001, 0.01, 0.1]
- Max iterations: [1000, 2000, 5000]
- Initial learning rate (eta0): [0.001, 0.01, 0.1]

Validation Strategy:

- 5-fold cross-validation for hyperparameter selection
- Multi-seed validation (seeds: 42, 123, 456, 789, 2024) for robustness assessment
- All four features used for maximum predictive accuracy
- Standardized features to ensure equal gradient contribution

3.2 OLS Regression Approach

The OLS implementation uses statsmodels for comprehensive statistical diagnostics:

Feature Set Candidates: Based on preprocessing analysis, five trustworthy feature combinations were selected:

- High Performance: Temp + Vacuum + Humidity ($R^2 \approx 0.928$)
- Most Stable: Temp + Pressure + Humidity ($R^2 \approx 0.921$)
- Simple Strong: Temp + Humidity ($R^2 \approx 0.920$)
- Balanced: Temp + Vacuum + Pressure ($R^2 \approx 0.918$)
- Temp-Vacuum: Temp + Vacuum ($R^2 \approx 0.915$)

Statistical Validation:

- Variance Inflation Factor (VIF) < 5 for multicollinearity assessment
- Condition number < 30 for numerical stability
- All coefficients significant ($p < 0.05$)
- Residual analysis: Jarque-Bera, Breusch-Pagan, Durbin-Watson tests
- Multi-seed validation for coefficient stability assessment

3.3 Model Comparison Framework

Both models will be evaluated using:

- R^2 (coefficient of determination)
- RMSE (root mean squared error)
- MAE (mean absolute error)
- Cross-validation stability
- Statistical significance (OLS)
- Hyperparameter sensitivity (SGD)

4. Expected Results and Interpretation

Based on preliminary analysis, we anticipate:

- Both models achieving $R^2 > 0.92$ on test data
- Strong negative coefficients for temperature and vacuum
- Positive coefficients for pressure and humidity
- SGD showing sensitivity to learning rate and regularization
- OLS providing interpretable coefficients with statistical significance
- Temperature emerging as the most influential predictor across both approaches

The analysis will provide insights into combined cycle power plant efficiency and the relative importance of ambient conditions on electrical energy output.