Project 1 Final Report

Introduction

Concrete is a fundamental material in construction, and its compressive strength is critical to ensuring structural integrity and longevity. The ability to predict and optimize compressive strength based on the composition of concrete mixtures is crucial for engineering high-performance materials that are cost-effective and environmentally sustainable.

Concrete strength depends on several predictor variables, including the proportion of cement, water, aggregates (fine and coarse), and additives like fly ash, blast furnace slag, and superplasticizers. Each of these components interact in complex ways, affecting the final strength of hardened concrete.

The primary motivation behind analyzing this dataset is to quantify the relationships between these predictor variables and compressive strength using statistical and machine learning techniques. By applying univariate and multivariate linear regression models, we can determine which components have the most significant impact on strength and assess whether linear relationships adequately capture these effects.

This study employs two approaches:

1. Gradient Descent-Based Linear Regression: A self-coded optimization algorithm used to iteratively adjust model parameters for best performance.
2. Ordinary Least Squares (OLS) Regression: A widely used statistical approach that estimates regression coefficients and provides statistical significance metrics such as p-values

By comparing these models, we aim to determine the most influential factors in concrete strength and provide actionable recommendations for optimizing concrete mixtures. Additionally, by evaluating model performance on both training and testing data, we assess our models and identify potential limitations in predictive accuracy.

The broader implications of this study extend beyond academic interest. If engineers can reliably predict and optimize concrete strength based on material composition, they can reduce material costs, improve sustainability, and enhance safety/durability.

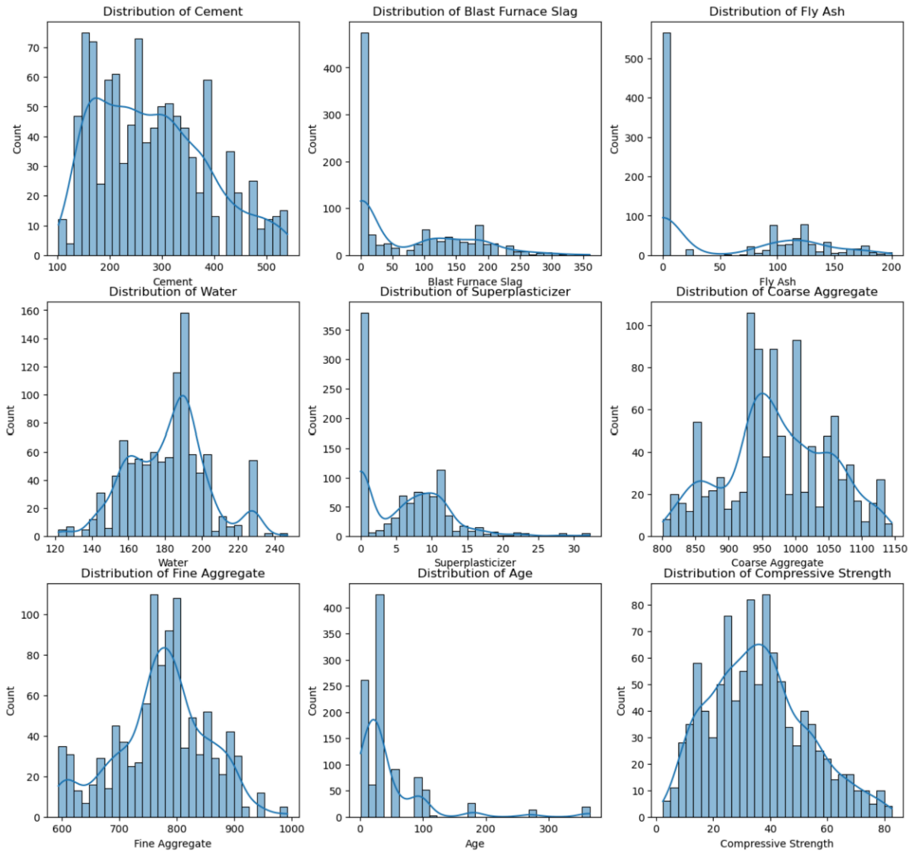
Results

The concrete dataset used in this study includes 1030 instances and 9 attributes (8 input variables and 1 output variable). See table 1 for variable details and table 2 for variable distributions:

Table 1. Dataset

Variable	Type	Units
Cement	Continuous	kg/m^3
Blast Furnace Slag	Continuous	kg/m^3
Fly Ash	Continuous	kg/m^3
Water	Continuous	kg/m^3
Superplasticizer	Continuous	kg/m^3
Coarse Aggregate	Continuous	kg/m^3
Fine Aggregate	Continuous	kg/m^3
Age	Discrete	Day
Concrete Compressive Strength (target)	Continuous	MPa

Table 2. Distributions



Univariate Linear Model (Part B)

In Part B we implemented univariate linear regression models using gradient descent to evaluate the relationship between each individual predictor variable and compressive strength. The goal was to determine which single factors have the strongest predictive power in estimating concrete strength.

1. **Model Performance and Key Findings**

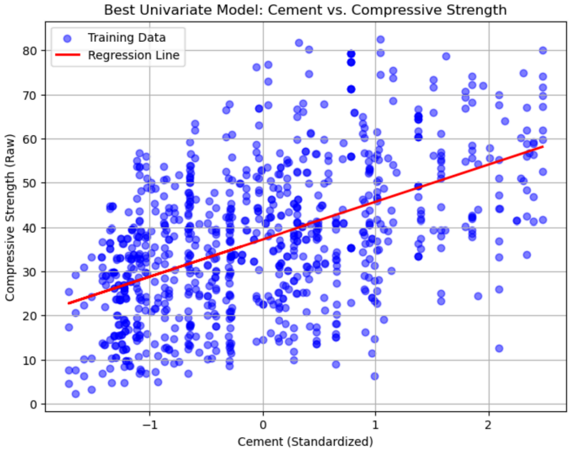
   Each predictor variable was tested separately to fit a linear regression model using gradient descent. We experimented with different learning rates and stopping conditions to optimize convergence. The models were evaluated based on Mean Squared Error (MSE) and Variance Explained (VE)

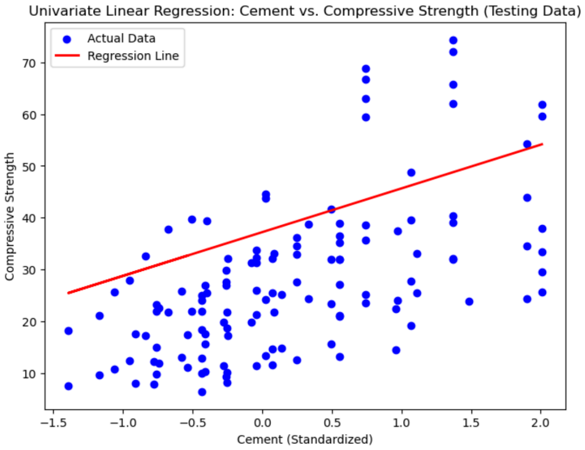
   Below are the final parameter values and performance metrics for the univariate models when using *standardized* values for the predictors and raw values for the response:

Predictor	Slope (m)	Intercept (b)	MSE	VE
Cement	8.4383	37.2146	203.5274	0.2655
Blast Furnace Slag	2.6919	36.9019	270.2041	0.0248
Fly Ash	-3.4109	37.1494	265.3488	0.0423
Water	-4.3946	36.8017	256.0473	0.0759
Superplasticizer	5.2727	36.3856	249.2388	0.1005
Coarse Aggregate	-2.1332	36.9861	272.5492	0.0164
Fine Aggregate	-2.4545	36.9291	270.7920	0.0227
Age	5.9977	36.9357	243.1905	0.1223

2. **Interpretation of Results**

   a. Cement is the most important predictor: The Cement predictor had the highest VE value (0.2655), indicating that cement content alone explains approximately 26.55% of the variance in compressive strength. Below are visuals on the training and testing data:





b. Fly Ash, Coarse Aggregate, Fine Aggregate and Water have a negative impact on strength. Specifically, the Water predictor had the most negative coefficient (m = -4.3946), meaning that as water content increases, compressive strength decreases.

c. Variables like Fly Ash, Blast Furnace Slag, Coarse Aggregate, and Fine Aggregate had low VE values (< 0.05), suggesting that they have weak individual correlations with compressive strength when analyzed in isolation.

3. Why These Results Can Be Trusted

a. Gradient Descent Optimization: We fine-tuned learning rates and stopping conditions to ensure convergence

b. Standardized Predictors: This ensured that variables with larger scales (ex: Age) did not dominate those with smaller magnitudes

c. Training and Testing Separation: The models were trained on 900 samples and tested on 130 samples to reduce overfitting

d. Consistency with Engineering Knowledge: The negative impact of water and the positive impact of cement align with well-known concrete science principles

4. Recommendations Based on These Results

a. Increase Cement Content: Since cement is the strongest predictor, optimizing its proportion in the mix is crucial

b. Limit Water Usage: Given its negative impact, minimizing water while incorporating superplasticizers to maintain workability is recommended

c. Consider Multivariate Effects: Since some variables had weak univariate relationships, a multivariate model (explored in Part C) may provide a more comprehensive understanding

Multivariate Linear Model (Part C)

In Part C we built a multivariate linear regression model using all 8 predictor variables to analyze their combined impact on concrete compressive strength. Unlike the univariate models, this approach accounts for interactions between features, leading to a more robust understanding of how multiple factors contribute to concrete strength.

For this analysis, we pre-processed the data by applying a log transformation to Cement, Age, Superplasticizer and Water before fitting the model. This was done to improve the statistical significance of certain variables by making their distributions more normal, reducing the effect of outliers and improving model interpretability. The response variable remained in its raw form.

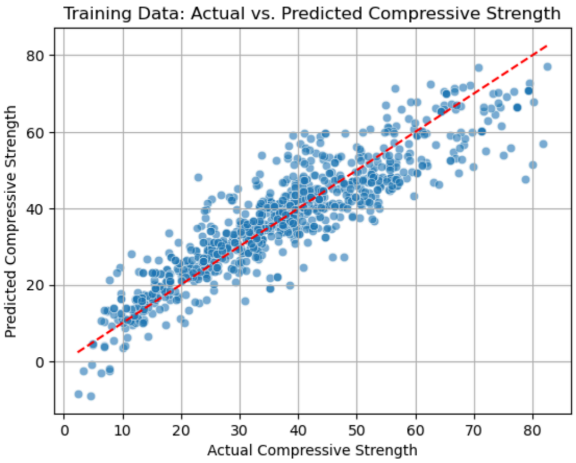
1. **Model Performance and Key Findings**

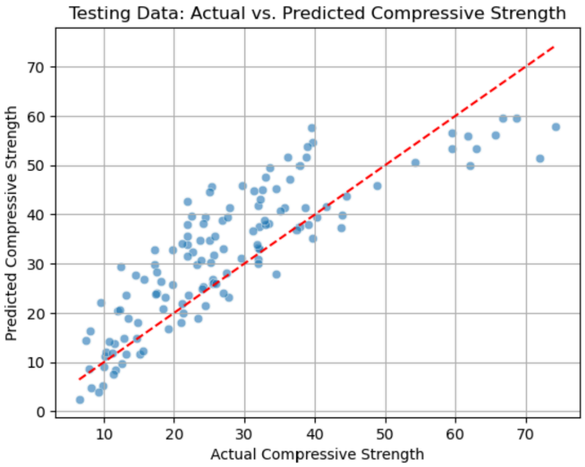
   The multivariate linear regression model was trained using Ordinary Least Squares (OLS) from the statsmodel package, which allowed us to obtain parameter estimates, p-values, and model statistics. Below are the key results:

Predictor	Coefficient (m)	p-value
Cement	34.6351	3.147857e-88
Blast Furnace Slag	0.1168	3.285967e-53
Fly Ash	0.0740	1.349350e-14
Water	-25.9490	5.186788e-10
Superplasticizer	0.9645	1.124314e-02
Coarse Aggregate	0.0287	4.757934e-07
Fine Aggregate	0.0278	1.888529e-05
Age	9.3214	1.552608e-216

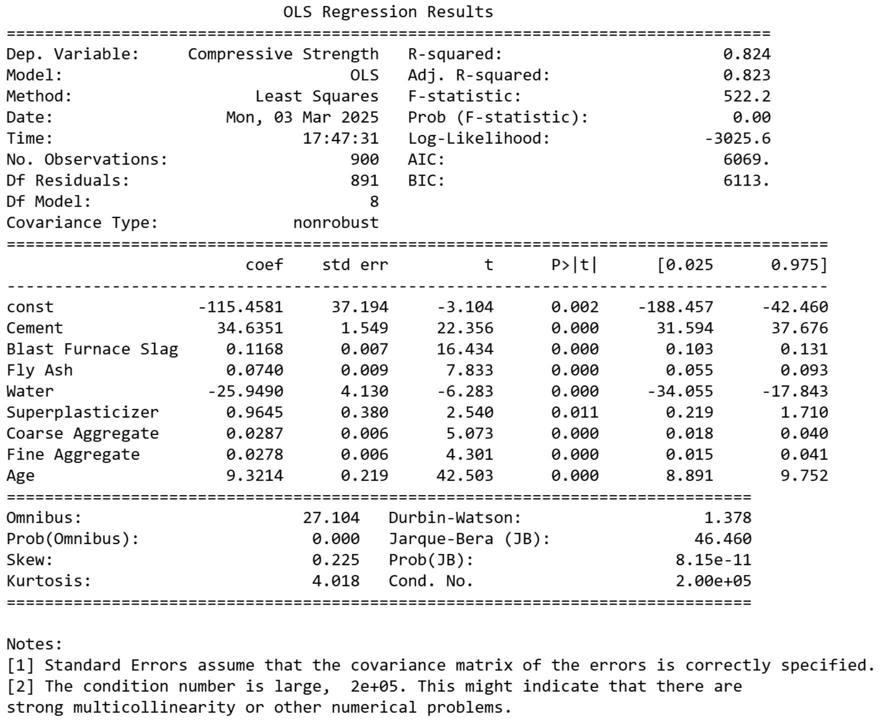
Training Variance Explained	0.824223
Testing Variance Explained	0.625056

Below are visualizations on training and testing data:





Below is the output of the OLS Regression model:



```
                             OLS Regression Results
================================================================================
Dep. Variable:     Compressive Strength   R-squared:                       0.824
Model:                              OLS   Adj. R-squared:                  0.823
Method:                   Least Squares   F-statistic:                     522.2
Date:                  Mon, 03 Mar 2025   Prob (F-statistic):               0.00
Time:                          17:47:31   Log-Likelihood:                -3025.6
No. Observations:                   900   AIC:                             6069.
Df Residuals:                       891   BIC:                             6113.
Df Model:                             8
Covariance Type:              nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const               -115.4581     37.194     -3.104      0.002    -188.457     -42.460
Cement                34.6351      1.549     22.356      0.000      31.594      37.676
Blast Furnace Slag     0.1168      0.007     16.434      0.000       0.103       0.131
Fly Ash                0.0740      0.009      7.833      0.000       0.055       0.093
Water                -25.9490      4.130     -6.283      0.000     -34.055     -17.843
Superplasticizer       0.9645      0.380      2.540      0.011       0.219       1.710
Coarse Aggregate       0.0287      0.006      5.073      0.000       0.018       0.040
Fine Aggregate         0.0278      0.006      4.301      0.000       0.015       0.041
Age                    9.3214      0.219     42.503      0.000       8.891       9.752
==============================================================================
Omnibus:                       27.104   Durbin-Watson:                   1.378
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               46.460
Skew:                           0.225   Prob(JB):                     8.15e-11
Kurtosis:                       4.018   Cond. No.                     2.00e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large,  2e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

2. Interpretation of Results

a. Multivariate Model Outperforms Univariate Model: Compared to the univariate models in Part B, which had the highest VE = 0.2655 (Cement), the multivariate model achieved a VE of 0.8242 on the training data. This demonstrated that considering multiple predictors together significantly improves predictive power.

b. Age is the most significant predictor (p < 0.0001), confirming that concrete strength increases over time.
c. Water Has a Negative Impact: The negative coefficient for Water (-25.9490) aligns with the fact that excess water weakens concrete strength. The p-value (p < 0.0001) confirms this effect is statistically significant.
d. Superplasticizer has the least impact on strength (p ≈ 0.01).
e. Many predictors that were weak individually (ex: Blast Furnace Slag) became stronger when combined

3. **Why These Results Can Be Trusted**

    a. Statistical Significance: p-values indicate which predictors truly affect compressive strength, helping eliminate noise
    b. Data Pre-Processing: Log transformation reduced skewness and improved the normality of features, enhancing model reliability
    c. Multivariate Effects: Unlike univariate models, this approach captures how multiple factors interact in real-world concrete mixtures

4. **Recommendations Based on These Results**

    a. Optimize Cement Usage: Since Cement has the second highest impact, maximizing it (within cost constraints) will yield the greatest strength improvements.
    b. Control Water Content: Since Water negatively affects strength, reducing it while using superplasticizers strategically may improve results
    c. Leverage Age-Strength Relationship: Builders should test concrete strength at different curing times to optimize long-term performance

Methods

Univariate Linear Regression (Part B):

1. **Objective**: Train separate linear models for each predictor variable to understand their individual contributions to concrete compressive strength.
2. **Pre-processing**: I applied standardization to predictor variables to have a mean of 0 and a standard deviation of 1, but kept the response variable (Compressive Strength) in its raw form. This approach helps ensure numerical stability during gradient updates. The predictor variables were standardized using StandardScaler from sklearn.preprocessing
3. **Hyperparameter Selection**: Learning rates were manually tuned (through trial and error) for each predictor to avoid overshooting the minimum while ensuring convergence within a reasonable number of iterations. Learning rates were adjusted until at least two predictors had positive variance explained. Lower values were used for features that exhibited slower convergence. See table below for each learning rate:

Predictor	Learning Rate
Cement	0.1
Blast Furnace Slag	0.02
Fly Ash	0.02
Water	0.01
Superplasticizer	0.05
Coarse Aggregate	0.02
Fine Aggregate	0.02
Age	0.02

4. **Dataset Splitting**: The dataset was split into:
    a. Training set: Rows 0-500 and 631-1030
    b. Testing set: Rows 501-630
5. **Fitting Process**: The gradient descent algorithm was implemented as follows:
    a. Initialize parameters: m = 1; b = 1
    b. Learning rate, alpha, was tuned separately for each predictor
    c. The maximum number of iterations was set to 40,000 to ensure convergence
    d. Predictions were computed as: $\hat{y} = mX + b$
    e. The error was calculated as: error = $y - \hat{y}$
    f. Gradients were computed using the partial derivatives of MSE
    g. Parameters were updated by subtracting alpha times its gradient from its current value
6. **Performance Metrics**: Mean Squared Error was computed to evaluate model performance. Variance Explained was used to assess how much of the variance in the target variable was captured.

Multivariate Linear Regression (Part C):

1. **Objective**: Train a model that considers all eight predictors simultaneously to determine their combined effect on concrete strength.
2. **Pre-processing**: I applied log transformation to Cement, Age, Water, and Superplasticizer to reduce the impact of extreme values, but kept the response variable (Compressive Strength) in its raw form. The transformation was done using: $X = \log(1+X)$.
3. **Dataset Splitting**: The dataset was split into:
    a. Training set: Rows 0-500 and 631-1030
    b. Testing set: Rows 501-630
4. **Adding a Constant Term**: To include an intercept term (bias) in the regression model, a column of ones was added to both the training and testing data

5. **Fitting Process**: The model was fitted using the Ordinary Least Squares (OLS) regression provided by the statsmodel package. The regression coefficients were estimated by minimizing the sum of squared residuals
6. **Statistical Testing**: p-values were computed for each predictor using the t-statistic, indicating feature significance. Lower p-values (<0.05) suggest significant evidence that a predictor affects compressive strength.
7. **Performance Metrics**: Variance Explained was used to assess how much of the variance in the target variable was captured. Scatter plots of actual vs. predicted values were also generated for both training and testing sets to assess accuracy.

Conclusion

The insights from this study suggest that optimizing concrete strength relies primarily on factors such as Cement content and curing Age, while other materials, like Aggregates, have a weaker influence. A practical recommendation would be to focus on optimizing Cement content and the curing process to improve strength efficiently. Additionally, incorporating nonlinear interactions or higher-order regression models may further improve predictive performance.

This study demonstrates the power of data-driven insights in material science, where statistical modeling and machine learning can optimize engineering processes.