TU Dortmund

Introductory Case Studies

Project 3: Regression Analysis

Lecturers:

Prof. Dr. Katja Ickstadt

M. Sc. Zeyu Ding

M. Sc. Marvin Napps

Author: Sadat Mahmud

Group number: 2

Group members: Sadat Mahmud, Ahmet Emircan Coskun, Montasir Hasan Chowdhury, Manouchehr Norouzi

January 26, 2024

Contents

1 Introduction

The study of the correlation between the composition of concrete and its compressive strength is a crucial endeavor in the field of civil engineering since the compressive strength of concrete is a vital metric. Concrete's compressive strength is its ability to support weights before breaking. The compressive strength test is the most crucial of the various tests conducted on the concrete since it provides insight into the material's properties (Jaya (2020)). The main goal of this analysis is to look at the connections between various elements that are important to the composition of concrete. Various quantitatively independent components are used in the experiment, such as cement, fly ash, blast furnace slag, water, superplasticizer, coarse aggregate, and fine aggregate. The goal is to comprehend the complex interactions among these concrete components and how they affect compressive strength over time.

The central objective of this assignment is to conduct a comprehensive regression analysis to unravel how distinct ingredients, alongside the age of concrete, contribute to its compressive strength. Recognizing the highly nonlinear nature of concrete strength, influenced by both components and maturation time, the anticipated outcome of this analysis is the development of a predictive model. This model, once established, holds the potential to inform the formulation of more effective and durable concrete mixtures. Beyond its academic significance, the insights gained from this study could have profound practical implications in shaping the approach to formulating concrete mixtures for diverse applications in civil engineering.

Section 2 provides a thorough description of the dataset and the project's goals.

A variety of statistical techniques are covered in Section 3, including the creation and evaluation of statistical models, the standards for selecting models, statistical tests to determine the significance of parameters, the assessment of goodness of fit, and the application of the Variance Inflation Factor (VIF) to examine multicollinearity, residual analysis.

Section 4 presents the previously specified statistical methods that are applied to the given dataset, and then the results are examined and interpreted.

In Section 5, the findings are presented along with a thorough project summary. The section ends with discussing possible directions for this project's future research.

2 Problem statement

2.1 Overview of Dataset

The "Introductory Case Studies" course instructors at Technische Universität Dortmund during the winter of 2023/2024 provided the dataset used in this study. The data was sourced from a website "UC Irvine Machine Learning Repository" (uci). The machine learning community uses the UCI Machine Learning Repository as a repository for databases, domain theories, and data generators to empirically analyze machine learning algorithms (uci). The dataset under investigation has nine variables and 1,030 observations. Eight quantitative independent variables are present in this set: cement, fly ash, blast furnace slag, water, superplasticizer, coarse aggregate, and fine aggregate. The dataset also contains the compressive strength, which is the dependent variable. According to a preliminary analysis, all variables in the dataset have no missing values. The absence of missing values simplifies the data preprocessing step and ensures we can analyze all datasets without resorting to imputation or other methods for handling missing data.

2.2 Project Objective

The main objective of this study is to use rigorous regression analysis techniques to improve our understanding of the correlations between compressive strength and concrete composition. Conduct descriptive analyses, specifically by generating scatter plots or correlation plots, to offer a succinct overview of the relationships between the variables in the dataset. This aims to lay the groundwork for a comprehensive understanding of the interdependencies among the considered variables. Consequently, Create a linear regression model for the compressive strength of concrete, including all the provided factors. Using polynomial regression factors, examine any non-linear connections explicitly involving the variable 'Age.' Select appropriate explanatory variables for concrete compressive strength using criteria like AIC, BIC, adjusted-R^2, or Mallows' Cp values, summarize regression results, and provide thorough interpretation. Thus, employ the chosen model to generate residual plots for model evaluation. Examine these plots to explore the presence of linear patterns, heteroskedasticity, and normal distribution. In addition, evaluate multicollinearity by utilizing the variance inflation factor (VIF). Ultimately, utilize the subset of identified explanatory variables to construct the ultimate

regression model. Engage in a concise analysis of the possible obstacles or restrictions linked to the ultimate model, providing valuable perspectives for future enhancements.

3 Statistical methods

This chapter comprehensively covers all the statistical methodologies utilized in this study. We applied the Python programming language (Rossum and Drake (2009)) in the Jupyter Notebook computational environment for our analysis. We chose to integrate essential libraries and packages, including Pandas (McKinney (2010)), Seaborn (Waskom (2012)), Matplotlib (Hunter (2007)), Scipy (Eric Jones and Travis Oliphant and Pearu Peterson and others (2001)), Itertools (Itertools Software Foundation (2003) into our Jupyter Notebook environment. This decision enabled the process of interactive development and comprehensive code documentation.

3.1 Linear Regression

In this project, regression analysis serves as a fundamental tool for modeling the relationship between the response variable y and a set of explanatory variables $x_1, \ldots, x_k$. Regression analysis is a statistical methodology that characterizes the connection between a dependent variable and one or more independent variables (Fahrmeir et al. (2013) p. 5-6). The goal is to estimate the unknown function f, which represents the systematic relationship between y and $x_1, \ldots, x_k$, while considering the impact of random noise denoted as ϵ (Fahrmeir et al. (2013) p. 73-75). The linear model is expressed as $y = f(\mathbf{x}) + \epsilon$, with $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$, where $\beta_0, \beta_1, \ldots, \beta_k$ are the parameters to be estimated (Fahrmeir et al. (2013) p. 73-75).

The assumptions made within the linear model include the linearity of the systematic component f, where $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. The intercept β_0 accounts for scenarios where all covariates are zero. Despite the initial appearance of linearity being restrictive, the model can effectively capture nonlinear relationships within the framework of linear models (Fahrmeir et al. (2013) p. 73-75).

Another essential assumption is the presence of additive errors, implying that the observed response y is the sum of the systematic component $f(\mathbf{x})$ and the random noise ϵ. The overall model equation is $y = \mathbf{X}\beta + \epsilon$, where $\mathbf{X}$ is the design matrix comprising intercept and covariate values (Fahrmeir et al. (2013) p. 73-75).

To estimate the unknown parameters β, data pairs $(y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$ are collected. Each observation is represented by the equation $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$, where ϵ_i is the error term for the i-th observation (Fahrmeir et al. (2013) p. 73-75).

The design matrix $\mathbf{X}$ is defined as:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}$$

It includes a column of ones for the intercept and the values of covariates for each observation. It is crucial for $\mathbf{X}$ to have full column rank to ensure unique estimators of the regression coefficients (Fahrmeir et al. (2013) p. 73-75).

Assumptions about the errors ϵ_i include having an expectation of zero, constant error variance σ^2, and homoscedastic and uncorrelated errors, which lead to a diagonal covariance matrix (Fahrmeir et al. (2013) p. 73-75).

In addition to the linear model, polynomial regression is considered in this statistical method. Polynomial regression extends the linear model to capture more complex relationships by introducing higher-degree polynomials, allowing for a better fit to the data.

Polynomial regression is a type of regression analysis in statistics that extends the linear regression model to accommodate relationships between variables that are nonlinear. While linear regression models relationships as a straight line, polynomial regression fits a polynomial equation to the data, allowing for a curved relationship between the independent and dependent variables (Fahrmeir et al. (2013) p. 139-140). The general form of a polynomial regression equation of degree n is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \epsilon$$

Here, y is the dependent variable, x is the independent variable, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients to be estimated, n is the degree of the polynomial, ϵ represents the error term (Fahrmeir et al. (2013) p. 139-140). The polynomial equation includes terms of various powers of x, up to the n-th power. The choice of the polynomial degree (n) depends on the complexity of the underlying relationship in the data (Fahrmeir et al. (2013) p. 139-140).

3.2 Akaike information criterion (AIC)

The Akaike Information Criterion (AIC) is a crucial tool for model selection within the framework of likelihood-based inference (Fahrmeir et al. (2013) p. 148). AIC serves as a quantitative measure to assess and compare the appropriateness of different models based on their fit to the observed data. The formulation of AIC, denoted as AIC, is given by:

$$\text{AIC} = -2 \cdot \mathcal{L}(\hat{\boldsymbol{\theta}}_{\text{OM}}, \hat{\sigma}^2) + 2 \cdot p,$$

Where $\mathcal{L}(\hat{\boldsymbol{\theta}}_{\text{OM}}, \hat{\sigma}^2)$ represents the maximum log-likelihood achieved when the Maximum Likelihood (ML) estimators $\hat{\boldsymbol{\theta}}_{\text{OM}}$ and $\hat{\sigma}^2$ are substituted into the log-likelihood function. The term p corresponds to the total number of parameters in the model, including the error variance σ^2 (Fahrmeir et al. (2013) p. 148).

For a linear model with Gaussian errors, the expression for AIC is further specified as:

$$\text{AIC} = n \cdot \log(\hat{\sigma}^2) + 2 \cdot (p + 1),$$

Where n is the number of observations. It is noteworthy that the ML estimator $\hat{\sigma}^2$ is employed in AIC rather than the conventional unbiased variance estimator (Fahrmeir et al. (2013) p. 148).

The given formulation uses AIC to measure the balance between the model's goodness of fit and complexity, penalizing excessive model complexity. Smaller AIC values indicate a better balance between fitting the data and avoiding overfitting. This criterion is particularly valuable in our project for selecting models that effectively capture the underlying patterns in the dataset while avoiding unnecessary complexity (Fahrmeir et al. (2013) p. 148).

3.3 Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) is a criterion commonly used for model selection within the context of likelihood-based inference (Fahrmeir et al. (2013) p. 149-150). BIC is generally defined as:

$$\text{BIC} = -2 \cdot \mathcal{L}(\hat{\boldsymbol{\theta}}_{\text{OM}}, \hat{\sigma}^2) + \log(n) \cdot (|\mathcal{M}| + 1),$$

Where $\mathcal{L}(\hat{\boldsymbol{\theta}}_{\text{OM}}, \hat{\sigma}^2)$ represents the maximum log-likelihood achieved when the Maximum Likelihood (ML) estimators $\hat{\boldsymbol{\theta}}_{\text{OM}}$ and $\hat{\sigma}^2$ are substituted into the log-likelihood function (Fahrmeir et al. (2013) p. 149-150). The term $|\mathcal{M}|$ denotes the total number of parameters in the model, and n is the number of observations.

For a linear model with Gaussian errors, the BIC can be expressed as:

$$\text{BIC} = n \cdot \log(\hat{\sigma}^2) + \log(n) \cdot (|\mathcal{M}| + 1).$$

Here, N is the number of datasets, $\log(\hat{\sigma}^2)$ is the natural logarithm of the estimated error variance. $\hat{\sigma}^2$ is the Maximum Likelihood (ML) estimator for the error variance. $\log(n)$ is the natural logarithm of the number of observations. $|\mathcal{M}|$ is the cardinality of the model space, representing the total number of parameters in the model. $+1$ is a penalty term to account for the number of parameters, preventing overfitting (Fahrmeir et al. (2013) p. 149-150).

The BIC is similar in form to the Akaike Information Criterion (AIC), but it penalizes complex models more strongly. Smaller BIC values indicate a better model fit. It's important to note that BIC and AIC are motivated differently, with BIC favoring more parsimonious models (Fahrmeir et al. (2013) p. 149-150).

In practical terms, the BIC is particularly useful for selecting models that balance goodness of fit with model simplicity, with a stronger preference for simplicity compared to AIC (Fahrmeir et al. (2013) p. 149-150).

3.4 Confidence Interval

The duality between two-sided tests and confidence intervals or confidence regions allows us to construct a confidence interval for a single parameter β_j, where $j = 0, \ldots, k$, or a confidence ellipsoid for a subvector β_1 of β. To construct a confidence interval for β_j under normality, we utilize the test statistic $t_j = \frac{\hat{\beta}_j - \beta_j}{\text{se}_j}$ corresponding to the test $H_0 : \beta_j = \beta_j^*$. Rejection of the null hypothesis occurs when $|t_j| > t_{n-p-1,1-\frac{\alpha}{2}}$. This test is designed such that the probability of rejecting H_0 when H_0 is true equals α. Therefore,

under H_0, we have

$$P(|t_j| > t_{n-p-1,1-\frac{\alpha}{2}}) = \alpha.$$

Consequently, the probability of not rejecting H_0 (given H_0 is true) is expressed as

$$P(|t_j| < t_{n-p-1,1-\frac{\alpha}{2}}) = P\left(\left|\frac{\hat{\beta}_j - \beta_j}{\text{se}_j}\right| < t_{n-p-1,1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

This leads to the following inequality:

$$P\left(\hat{\beta}_j - t_{n-p-1,1-\frac{\alpha}{2}} \cdot \text{se}_j < \beta_j < \hat{\beta}_j + t_{n-p-1,1-\frac{\alpha}{2}} \cdot \text{se}_j\right) = 1 - \alpha,$$

resulting in a $(1-\alpha)$-confidence interval for β_j:

$$\left[\hat{\beta}_j - t_{n-p-1,1-\frac{\alpha}{2}} \cdot \text{se}_j, \hat{\beta}_j + t_{n-p-1,1-\frac{\alpha}{2}} \cdot \text{se}_j\right].$$

In a similar way, a $(1-\alpha)$-confidence region can be constructed for an r-dimensional subvector β_1 of β (Fahrmeir et al. (2013) p. 136-137). This approach provides a robust measure for estimating parameters in this project, ensuring a proper balance between fitting the model and avoiding overfitting.

3.5 Best Subset Selection

In this project, best subset selection is employed to identify the optimal predictors for the linear regression model. This method systematically explores all possible combinations of the p, predictors, fitting separate least squares regressions for each combination (James et al. (2013) p. 205-207). The process begins with the null model (M_0) predicting the sample mean. Subsequently, for each subset size (k), all $2^{(p-1)}$ models containing k predictors are evaluated (James et al. (2013) p. 205-207). The best model (M_k) in each subset size is determined based on the largest R-squared (R^2) (James et al. (2013) p. 205-207). The final step involves selecting the single best model from M_0 to M_p, considering criteria such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or adjusted R^2 (James et al. (2013) p. 205-207). This meticulous approach mitigates the risk of overfitting, especially since the R^2 tends to decrease monotonically with the inclusion of more features. In addressing the computational challenges, it is acknowledged that the exhaustive consideration of 2^p models becomes

impractical for large values of p (James et al. (2013) p. 205-207). The computational efficiency of best subset selection is limited, particularly when p exceeds 40, necessitating alternative methods to handle such high-dimensional datasets effectively (James et al. (2013) p. 205-207).

3.6 Multicollinearity Analysis

Multicollinearity occurs when there is a significant connection among the independent variables in regression analysis. This is undesirable since it might result in imprecise estimations of regression coefficients. These phenomena are closely connected to the linear regression assumption that requires the design matrix to have a total rank. When multicollinearity is present, the inverse of the design matrix, represented as $(X'X)^{-1}$, does not exist, making the least squares method worthless. The extent of multicollinearity is measured using the variance formula for $\hat{\beta}_j$, where a more considerable variance value indicates a stronger linear relationship between the explanatory variable x_j and the other variables. The evaluation of the rise in $Var(\hat{\beta}_j)$ caused by this linear dependence is made more accessible by the variance inflation factor (VIF), which is calculated as

$$VIF_j = \frac{1}{1 - R_j^2}$$

The correlation coefficient R_j^2 represents the relationship between x_j and other variables. A high variance inflation factor (VIF) value for x_j suggests a higher level of collinearity. A Variance Inflation Factor (VIF) exceeding ten is widely considered indicative of collinearity problems, according to recommendations found in the relevant literature (Fox and Monette (1992) p.176-184). The Variance Inflation Factor (VIF) is employed when multiple coefficients are involved. To mitigate the issue of collinearity, a recommended strategy is to exclude the explanatory variables affected by the study. This approach aims to address the challenges posed by multicollinearity, enhancing the stability and reliability of regression analyses. Eliminating explanatory variables impacted by collinearity improves the precision of coefficient estimations and ensures the resilience of the statistical model. This methodology aligns with established principles in regression analysis, underscoring the significance of detecting and controlling for multicollinearity to yield accurate and relevant outcomes (Fahrmeir et al. (2013) p. 157-159).

3.7 Residuals Vs Fitted plot

In the assessment of heteroscedastic errors, it proves valuable to construct residuals versus fitted value plots, considering both the predicted values y_{Oi} and the covariates x_{ij}. Including covariates not incorporated in the model for a comprehensive analysis is crucial. Notably, plotting residuals against the response variable y itself is discouraged as residuals depend on the response variable, revealing this inherent dependency. To ensure a reliable examination of heteroscedasticity, standardized or studentized residuals are preferred over raw residuals, given the latter's intrinsic heteroscedasticity.

$$Var(\hat{e}_i) = \sigma^2(1 - h_{ij})$$

Consequently, raw residuals are deemed less suitable for heteroscedasticity assessment (Fahrmeir et al. (2013) p. 183). In instances of homoscedastic error variances, standardized or studentized residuals exhibit random fluctuations around zero with a consistent variance. Deviations from this pattern signal evidence of heteroscedastic variances (Fahrmeir et al. (2013) p. 183).

3.8 Q-Q plot

The Normal Probability Quantile-Quantile (Q-Q) plots offer a valuable means of comparing a sample distribution to the standard normal distribution. This graphical representation involves plotting sample quantiles (y) against theoretical quantiles (x) from the standard normal distribution. The procedure commences with sorting the sample in ascending order, resulting in observed quantiles $(y(1), y(2), ..., y(n))$. Plotting points (P_i) are computed using the formula:

$$P_i = \begin{cases} \frac{i-\frac{3}{8}}{n+\frac{1}{4}} & \text{if } n \leq 10 \\ \frac{i-\frac{1}{2}}{n} & \text{if } n > 10 \end{cases}$$

These points reveal outliers—observations deviating from the trend line formed by most data points. As evident in QQ-plots like qqData, outliers may suggest extreme values, prompting caution before considering their removal. Gaps in plotted points may signal issues such as rounding or a non-representative dataset. Thorough exploratory analysis is crucial, and outlier treatment should be approached with caution, considering potential data-gathering errors (Hay-Jahans p. 146-150).

4 Statistical Analysis

4.1 Correlation Patterns: A Descriptive Overview of Variable Associations

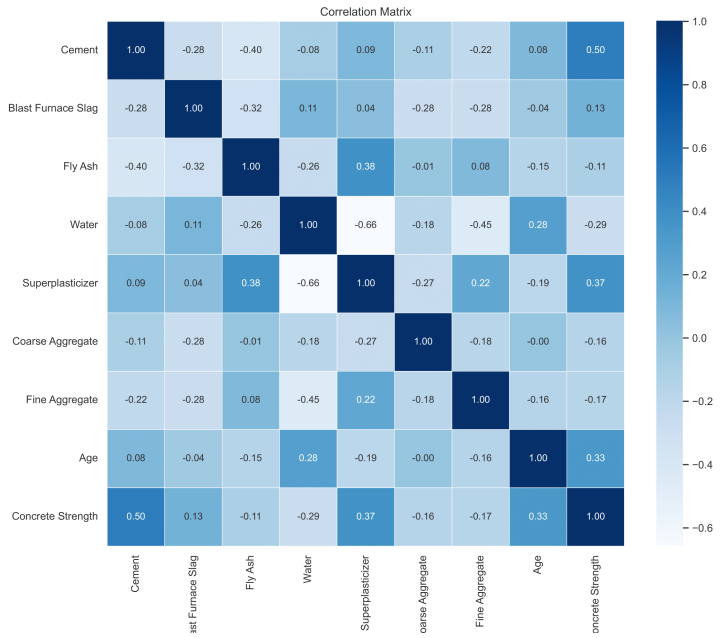

Figure 1: Correlation Matrix

The correlation matrix analysis uncovers substantial relationships among the variables being examined. Figure 1 is a correlation plot that visualizes the correlation values between variables. The strength of concrete shows a moderate positive connection with the amount of cement used (r = 0.50) and the level of superplasticizer added (r = 0.37). This suggests that higher cement content and enhanced incorporation of superplasticizers have a beneficial impact on the strength of concrete. In contrast, there is a modest positive correlation (r = 0.33) between the strength of the concrete and its age, indicating a subtle relationship between the age of the concrete mixture and its strength. Remarkably, there are negative relationships between concrete strength and water (r = -0.29), fine aggregate (r = -0.17), coarse aggregate (r = -0.16), and fly ash (r = -0.11). High water content, fine aggregate, coarse aggregate, and fly ash in concrete are linked to reduced concrete strength, as indicated by these negative correlations. The insights obtained by the correlation matrix provide a significant understanding of the

complex relationships between different components and their influence on the strength of concrete.

4.2 Polynomial Regression Model Overview

The task involved the implementation of a polynomial regression model to predict the concrete strength (y) utilizing various variables, with a specific focus on the age of the concrete. The regression model was constructed with a polynomial degree of 2, incorporating a quadratic component represented by Age^2.

Table 1: Regression Coefficients from Polynomial Regression Model

Variable	**Coefficient**
Intercept (const)	-3.60
Cement	0.11
Blast Furnace Slag	0.10
Fly Ash	0.07
Water	-0.17
Superplasticizer	0.20
Coarse Aggregate	0.01
Fine Aggregate	0.01
Age	0.35
Age^2	-0.00

Table 1 illustrates the R-squared value of the model, which is 0.743. This number indicates that the model can account for around 74.3% of the variation in concrete strength. The R-squared value has been modified to 0.741, considering the number of predictors in the model. Upon analyzing the statistical significance of each predictor, it is evident that all variables, except Coarse Aggregate and Fine Aggregate, exhibit p-values below 0.05, indicating their statistical importance. The contribution of Coarse Aggregate and Fine Aggregate to estimating concrete strength may be insignificant. The polynomial regression model also identifies possible concerns. The condition number is significantly high (1.73e+06), suggesting the presence of substantial multicollinearity or other numerical issues. It is crucial to exercise caution when interpreting coefficients and predictions in light of these issues. Although the polynomial regression model offers valuable insights into the correlation between variables and concrete strength, the concerns it reveals necessitate additional analysis and evaluation. Modifications made to

the model, aimed at addressing the problem of multicollinearity, can potentially improve its dependability.

4.3 Variable Subset Selection for Concrete Compressive Strength Prediction

Different model selection procedures were used to determine an appropriate subset of explanatory factors for predicting concrete compressive strength. The forward selection, backward selection, and best subset selection approaches were evaluated, guided by metrics such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and modified R-squared. Upon carefully evaluating these methodologies, the chosen subset of variables for the regression model comprises 'Cement,' 'Blast Furnace Slag,' 'Fly Ash,' 'Water,' 'Superplasticizer,' and 'Age.' The regression analysis yields a comprehensive summary of the coefficients, statistical significance (p-values), confidence intervals, and goodness-of-fit indices for the chosen variables, as presented in Table 2. The modified R-squared value of the model is 0.612, indicating that around 61.2% of the variation in concrete compressive strength can be accounted for by the selected selection of variables. The coefficients offer valuable insights into the correlation between each predictor variable and the strength of the concrete. For example, the variables 'Cement' and 'Age' correlate positively, whereas 'Water' has a negative effect. The statistical significance of these coefficients, as evidenced by the low p-values, underscores the dependability of the chosen subgroup in forecasting concrete strength. Nevertheless, acknowledging that a high condition number can indicate possible problems, such as multicollinearity, is essential.

Table 2: Regression Results for Selected Subset of Variables

Variable	**Coefficient**	**Std Err**	**t-value**	**P>|t|**	**[0.025**	**0.975]**
`const`	29.0302	4.212	6.891	0.000	20.764	37.296
`Cement`	0.1054	0.004	24.821	0.000	0.097	0.114
`Blast Furnace Slag`	0.0865	0.005	17.386	0.000	0.077	0.096
`Fly Ash`	0.0687	0.008	8.881	0.000	0.054	0.084
`Water`	-0.2183	0.021	-10.332	0.000	-0.260	-0.177
`Superplasticizer`	0.2390	0.085	2.826	0.005	0.073	0.405
`Age`	0.1135	0.005	20.987	0.000	0.103	0.124

4.4 Assessing Model Performance: Residual Analysis and Multicollinearity Check

The evaluation of the selected polynomial regression model in task 4 involved a comprehensive analysis of residual plots and diagnostic tests. Figure 2 illustrates the scatter plot of residuals against fitted values, serving as a critical assessment of linearity. The objective was to achieve a random scatter of points without discernible patterns, ensuring the model's adequacy.

Figure 2: Scale-Location Plot

In addition to the scatter plot, Figure 3 further contributes to our evaluation. It presents a detailed analysis of the histogram of residuals, demonstrating an approximate normal distribution. The accompanying Q-Q plot is a confirming indicator, providing assurance regarding the model's adherence to normal distribution assumptions.

The assessment of multicollinearity involved the application of the variance inflation factor (VIF). Our scrutiny revealed a minor concern, particularly with the constant term. This discovery emphasizes the need for careful consideration in maintaining the model's robustness. Addressing these nuances ensures the reliability and validity of the regression model.

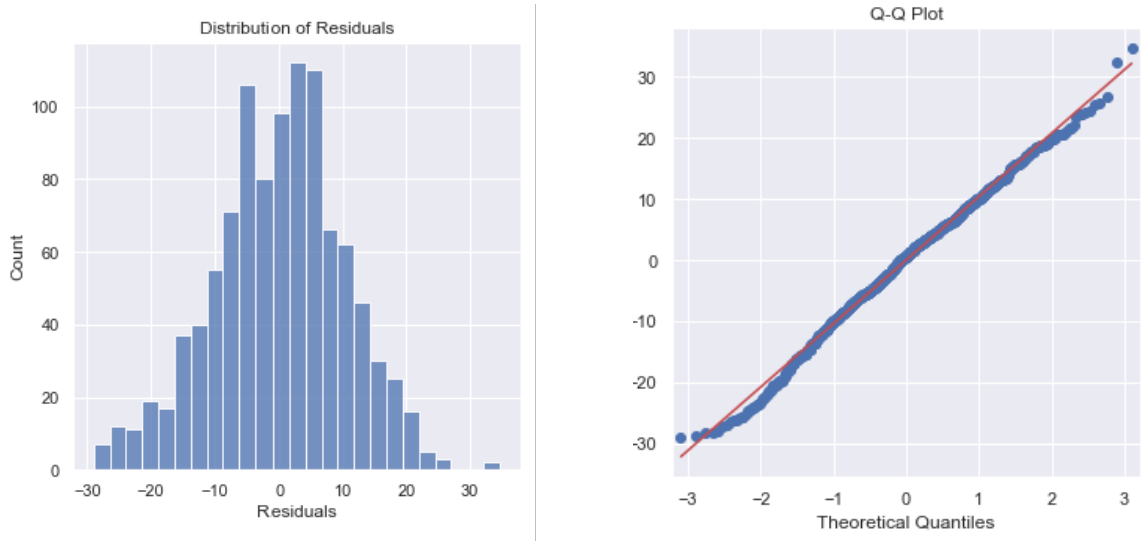

Figure 3: Normality of Residuals

4.5 Final Regression Model

As depicted in Table 2, the regression model delineates the relationship between Concrete Strength and the predictor variables Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, and Age, each associated with specific coefficients. The model's equation is expressed as

$$\begin{aligned}\text{Concrete Strength} = 29.0302 &+ 0.1054 \times \text{Cement} + 0.0865 \times \text{Blast Furnace Slag} \\ &+ 0.0687 \times \text{Fly Ash} - 0.2183 \times \text{Water} \\ &+ 0.2390 \times \text{Superplasticizer} + 0.1135 \times \text{Age}.\end{aligned}$$

Each coefficient signifies the change in Concrete Strength associated with a one-unit increase in the respective predictor while holding other variables constant. For instance, higher Cement, Blast Furnace Slag, Fly Ash, Superplasticizer, and Age correlate positively with increased Concrete Strength, while greater Water content is linked to a decrease. The statistical significance, as indicated by t-values and p-values, and the confidence intervals contribute to the comprehensive understanding of the model's reliability. However, it is essential to acknowledge potential issues, such as multicollinearity concerns highlighted by variance inflation factors, which may impact the model's robustness and merit further exploration and refinement.

5 Summary

The project aims to create a robust regression model for predicting concrete compressive strength, specifically for construction engineers and researchers. The dataset includes 1030 observations and nine predictor variables, including Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, Age, and an additional variable. The initial data analysis involved the analysis of a correlation matrix, followed by implementing a polynomial regression model with a quadratic component Age^2. The refined model, with predictors of Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age, accounted for approximately 61.2% of the variation in concrete compressive strength. Conducting residual analysis and examining multicollinearity using variance inflation factors (VIF) was crucial to ensure the model's reliability. The final regression model provides a mathematical formula for estimation. However, the high condition number raises concerns about potential multicollinearity issues, requiring further exploration and refinement. Future improvements could involve additional data collection and analysis to enhance the model's accuracy and applicability in real-world scenarios. give this text as latex overleaf text

Bibliography

UCI Machine Learning Repository: Concrete Compressive Strength. https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength. Accessed: [23/01/2024].

Eric Jones and Travis Oliphant and Pearu Peterson and others. SciPy Open Source Scientific Tools for Python, 2001. URL https://www.scipy.org/.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. The classical linear model. pages 73–75. Springer, 2013. URL https://doi.org/10.1007/978-3-642-34333-9$_3$.

John Fox and Georges Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 1992. doi: 10.1080/01621459.1992.10475190. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475190.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. Chapman Hall/CRC. ISBN 9781138329164.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Itertools Software Foundation. Python itertools Module, 2003. URL https://docs.python.org/3/library/itertools.html.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013. ISBN 978-1-4614-7137-0. with Applications in R.

Ramadhansyah Putra Jaya. Porous concrete pavement containing nanosilica from black rice husk ash. *Journal Name*, Volume Number:Page Range, July 2020. URL