TU Dortmund

Introductory Case Studies

Project 3: Regression Analysis

Lecturers:

Prof. Dr. Katja Ickstadt

M. Sc. Zeyu Ding

Author: Bilal Tanvir Bhatti

Group number: 1

Group members: Rafay Maqsood, Maimuna Rahman, Nafisa Farhin

January 25, 2024

Contents

1 Introduction

Studying how the ingredients in concrete affect its compressive strength is crucial in civil engineering. Concrete's strength is vital for its durability and how well it handles pressure in construction projects.

This investigation examines how different components in concrete influence its compressive strength, considering both linear and potential nonlinear connections. By analyzing how these ingredients interact, the goal is to understand what factors determine concrete's strength. This understanding can lead to stronger structural designs and better construction methods in civil engineering.

The main objective is to conduct a regression analysis to examine the impact of various factors. Initially, a descriptive analysis is performed on the dataset, then a correlation plot is used to identify patterns and associations between variables, indicating whether they tend to increase or decrease together, or if they are unrelated. Subsequently, a regression model is formulated, where Concrete compressive strength serves as the response variable, and other variables act as explanatory factors. Through the best subset selection technique, multiple models are evaluated, and the model with the lowest AIC and BIC values is chosen as the optimal model, featuring seven key features.

In Section 2, we offer an overview of the dataset and assess the data quality. Section 3 delves into the details and explanations of the statistical methods used for dataset analysis, including information on the formulas and assumptions of linear regression models. It also provides insights into the Akaike Information Criterion, Bayesian Information Criterion, and the coefficient of determination. Moving to Section 4, we discuss the pre-processing of the dataset and apply the methods outlined in Section 3 to analyze the dataset, interpreting the results. In conclusion, Section 5 briefly summarizes the study's results and discusses possible improvements for future experiments.

2 Problem Statement

2.1 Data set and Data Quality

The dataset being examined consists of $n = 1,030$ instances and encompasses 9 different factors. The data integrity is high, devoid of any absent values, indicating good overall quality. The dataset was compiled by the instructors of Introductory Case Studies at TU Dortmund and is available on `https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength`.

The entire dataset is utilized in the analysis, comprising 8 quantitative independent variables: *Cement* (1 decimal place, kilograms in a cubic meter mixture), *Blast Furnace Slag* (1 decimal place, kilograms in a cubic meter mixture), *Fly Ash* (1 decimal place, kilograms in a cubic meter mixture), *Water* (1 decimal place, kilograms in a cubic meter mixture), *Superplasticizer* (1 decimal place, kilograms in a cubic meter mixture), *Coarse Aggregate* (1 decimal place, kilograms in a cubic meter mixture), *Fine Aggregate* (1 decimal place, kilograms in a cubic meter mixture), and *Concrete compressive strength* (2 decimal places, MPa - megapascals) as the dependent variable.

2.2 Project Objectives

The main aim is to establish the best model for the given dataset, and specific tasks have been outlined to achieve this goal. The initial task involves utilizing a correlation plot to illustrate the connections between the variables. Subsequently, the project focuses on establishing a linear regression model for *Concrete Compressive strength* based on various independent variables: *Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Concrete Aggregate.* The third task involves identifying a suitable subset of explanatory variables for Concrete Compressive strength, determined by reduced AIC and BIC values. In the fourth task, residual plots are generated to evaluate the model, checking for linearity, heteroskedasticity, and normality patterns. Additionally, the presence of multicollinearity is assessed through the variance inflation factor (VIF), and any issues encountered in the final model are addressed and discussed.

3 Statistical methods

Several statistical methods that are used to analyze the data are discussed in this section. For this analysis, the R software (R Development Core Team, 2020), version 4.3.0 is used with packages ggplot2 (Wickham, 2016), leaps(based on Fortran code by Alan Miller, 2020), olsrr (Hebbali, 2020), readxl (Wickham and Bryan, 2023), corrplot (Wei and Simko, 2021), car (Fox and Weisberg, 2019).

3.1 Multiple Linear Regression

Multiple Linear Regression aims to depict how a set of k explanatory or independent variables, denoted as $x_1, ..., x_k$, influences a continuous target or response variable, y, by constructing a linear equation based on observed data. These predictors can encompass categorical or continuous variables. Unlike being a deterministic outcome of the covariates $f(x_1, ..., x_k)$, the response variable also incorporates random fluctuations or noise.

$$y_i = f(x_{1i}, ..., x_{ki}) = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \epsilon_i. \tag{1}$$

where x_{ij} represent the value of j^{th} covariate, $j = 1, , , , k$ for the i^{th} observation, $i = 1, ..., n$, ϵ_i represents the error term or the residual, which captures the discrepancy between the actual observed value y_i. (Fahrmeir et al., 2013, p. 74)

The undisclosed parameters $\beta_0, \beta_1, ..., \beta_k$ can be condensed into a $(k + 1)$-dimensional vector $\beta = (\beta_0, \beta_1, ..., \beta_k)$ where β_0 stands for the intercept. Hence,

$$y = f(\mathbf{x}) + \epsilon = \mathbf{x}'\beta + \epsilon.$$

To calculate the undisclosed parameters, data is gathered involving y_i and $x_i = (1, x_{i1}, ..., x_{ik})'$, where $i = 1, ..., n$. Consequently, the vectors $\mathbf{y}$ and ε are established as:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

and the design matrix $\mathbf{X}$ can be represented as:

$$\begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

3.1.1 Assumptions

The assumptions are the following:

- The errors have zero mean and can be written as $\mathbf{E}(\varepsilon) = 0$.
- The errors have constant variance among them and they are normally distributed.
- The design matrix $\mathbf{X}$ assumed to be full rank.

(Fahrmeir et al., 2013, p. 74-76)

Then n equations for Eq.1 can be summarized as:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon.$$

3.1.2 Parameter Estimation

The method of least squares stands out as the commonly utilized approach for parameter estimation in regression. This method assumes a linear relationship between independent and dependent variables. To differentiate between model parameters and their estimated values, the "hat" symbol ($\hat{\beta}_k$) is applied (Fahrmeir et al., 2013, p. 77). This distinction proves crucial as obtaining the true parameter value without any error is impractical. In consideration of this distinction, the estimator for the mean ($E(y_i)$) of the dependent variable (y_i) is expressed as:

$$E(y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \ldots + \hat{\beta}_k x_{i,k}$$

The estimated error, known as residuals, is defined as the variance between the actual value and the projected value.

$$\varepsilon_i = y_i - \hat{y}_i$$

The approach of least squares aims to minimize the total squared differences to estimate the unknown regression parameters (Fahrmeir et al., 2013, p. 105).

$$LS(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

The fair estimate of the β coefficients, under the condition that the design matrix X is linearly independent and $X'X$ is positive definite (Fahrmeir et al., 2013, p. 107), can be computed in this manner:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

3.1.3 Significance of Parameter Estimates

Typically, the assessment of parameter estimates significance is a common practice to gauge the statistical relevance of estimated parameters within the regression model. This evaluation aims to ascertain if the sample estimates of the parameters significantly deviate from zero. This scrutiny revolves around two main hypotheses: the null hypothesis (H_0) and the alternative hypothesis (H_1):

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned}$$

A p-value below a predetermined threshold, like 0.05, is frequently used to signify the statistical importance of the parameter estimate. This indicates that the association between the independent and dependent variables is statistically significant. To evaluate this significance, the t-test is commonly utilized, where the estimated parameter value is contrasted against zero through the t-distribution (Fahrmeir et al., 2013, p. 132). The test statistic employed under the null hypothesis is depicted as:

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var(\widehat{\beta_j})}}} \sim t_{n-p}$$

Here, j = (0, ..., k), n represents the count of observations, p stands for the size of the covariates vector, and $\sqrt{\widehat{Var(\widehat{\beta_j})}}$ represents the estimated standard deviation of the estimator $\hat{\beta}_j$.

3.1.4 Confidence Intervals on the Regression Coefficients

The uncertainty stemming from sampling has an impact on determining coefficients in a regression model. This uncertainty arises because the sample used for estimation is a random subset of the overall population, and the estimates are based on sample statistics. Confidence intervals, derived from sample data, provide a range of values wherein the parameter of a random variable's distribution can be expected to fall with a certain probability. These intervals are calculated using relevant sample statistics and a chosen confidence level, usually set at 95% (equivalent to 100(1 - α)), representing the probable range of values encompassing the population parameter.

Assuming a normal distribution of errors or a sufficiently large sample size (Fahrmeir et al., 2013, p. 137), the $(1 - \alpha)$ confidence interval for the estimate of $\hat{\beta}_j$ is expressed as:

$$[\hat{\beta}_j - t_{n-p}(1 - \frac{\alpha}{2}) \cdot \mathrm{se}_j, \hat{\beta}_j + t_{n-p}(1 - \frac{\alpha}{2}) \cdot \mathrm{se}_j]$$

The standard error se_j for the jth coefficient is determined utilizing the quantile level $(1 - \frac{\alpha}{2})$ of the t-distribution, where the degrees of freedom are $(n - p)$. In this context, n signifies the total number of observations, while p represents the length of the covariates vector.

3.2 Best Subset Selection

Best Subset Selection represents a method commonly applied in multiple linear regression scenarios, aimed at identifying a subset of p independent variables that most effectively clarify the outcome. Initially, the algorithm starts with a null model devoid of predictors (where $k = 0$), predicting the average mean for each observation. Subsequently, it progresses through all feasible subsets of independent variables, starting from $k = 1$ up to utilizing all available variables ($k = 1, 2, \ldots, p$). For each subset size, it fits models for all combinations using the least squares method, seeking the model with the smallest residual sum of squares. Eventually, the picked algorithm is the one, typically based on chosen criteria like the Akaike Information Criterion (AIC) or adjusted R^2 (Heiberger and Holland, 2015, p 639).

3.3 Akaike Information Criterion

The Akaike Information Criterion (AIC) serves as a statistical technique utilized in choosing models, assisting in comparing and identifying the most appropriate ones. To accommodate model complexity, AIC includes a penalty term, attributing higher scores to models featuring more parameters. The ideal model is chosen based on having the lowest AIC value. The computation of AIC is outlined as:

$$\text{AIC} \; = -2\log(\hat{L}) + 2k.$$

Here, k signifies the count of parameters essential for modeling and $\hat{L}$ stands for the maximum log-likelihood value within the model. This value serves as a determinant of how well the model fits the data. (Heiberger and Holland, 2015, p. 639)

3.4 Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is another statistical technique employed to pick the optimal model from a limited set of models. It resembles the Akaike Information Criterion (AIC), yet its primary distinction lies in penalizing more heavily complex models or those with an increased number of parameters compared to AIC. Like AIC, a lower BIC value for a model signifies a better fit. Mathematically, BIC can be expressed as:

$$\text{BIC} \; = -2\log(\hat{L}) + \log(n)k.$$

In this equation, k represents the count of parameters within a model and $\hat{L}$ stands for the maximum value attained by the likelihood function of that model.(Fahrmeir et al., 2013, p. 677-678)

3.5 R-Square

The coefficient of determination, often referred to as R^2, serves as a commonly used metric in regression analyses to evaluate the quality of fit by assessing the degree to which the model conforms to the data. A higher R^2 value ($R^2 = 1$) signifies a stronger fit, whereas a lower R^2 value ($R^2 = 0$) indicates a weaker fit. This metric quantifies how effectively the model captures the data by measuring the proportion of the dependent variable's variance explained by the model relative to the total variance in the dependent variable. (Fahrmeir et al., 2013, p 115) The formula for R^2 is given by:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

One drawback of R-squared (R^2) is its susceptibility to the number of independent variables in the model, as it does not decrease with the inclusion of additional variables. This characteristic poses challenges in comparing models using R-squared as a gauge of fit quality. (Fahrmeir et al., 2013, p 114)

A more reliable metric for comparing model fit is the adjusted R-squared (R^2 adjusted). The adjusted R^2 is calculated as follows:

$$\bar{R}^2 = 1 - \frac{(n-1)}{(n-p)}(1 - R^2)$$

Here, n signifies the count of observations, while p indicates the quantity of independent variables incorporated within the model.

3.6 Multicollinearity

Multicollinearity refers to a statistical scenario where there's a strong correlation between two or more predictor variables in a regression model. This correlation poses challenges in discerning the individual impact of each variable on the dependent variable. It can lead to unstable coefficient estimates, inflated standard errors, and difficulties in interpreting the significance and relevance of variables within the model.

An alternative method to evaluate multicollinearity involves calculating the variance inflation factor (VIF). The VIF quantifies the ratio between the variance of $\hat{\beta}_j$ from fitting the complete model and the variance of $\hat{\beta}_j$ if it were fitted independently. A VIF of 1 indicates no collinearity. Typically, some degree of collinearity exists among predictors. Generally, a VIF exceeding 10 indicates problematic collinearity. You can compute the VIF for each variable using the provided formula.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

In this context, $R^2_{X_j|X_{-j}}$ denotes the coefficient of determination derived from regressing the predictor X_j against all the other predictors. As $R^2_{X_j|X_{-j}}$ approaches one, it signals the existence of collinearity, resulting in a high value for the VIF. (James, 2013, p. 243)

3.7 Residual Plot

In regression analysis, residual plots serve as a tool to detect possible concerns regarding the linear relationship between the dependent and independent variables. An appropriately fitted regression model is expected to display residuals (the differences between actual and predicted values) evenly scattered around zero. Any noticeable patterns in the residuals may suggest issues with the linear model. In cases where nonlinear relationships exist in the data, applying methods like nonlinear transformations of predictors or integrating interaction terms into the model can prove advantageous.(James, 2013, p. 93)

4 Statistical analysis

In this section, the statistical techniques described earlier are applied to analyze the provided dataset and comprehend the outcomes.

4.1 Descriptive Analysis of the Variables

In this section, we are conducting a descriptive analysis of nine variables: Concrete compressive strength, Age, Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, and Fine Aggregate. Table.1 presents a summary of the data.

Table 1: Data summary

Variable name	**Min**	**1st Qu**	**Median**	**Mean**	**3rd Qu**	**Max**
Concrete-Compressive-Strength	2.33	23.71	34.44	35.82	46.13	82.59
Cement	102.00	192.40	272.90	281.20	350.00	540.00
Blast Furnace Flag	0.00	0.00	22.00	73.90	142.90	359.40
Fly Ash	0.00	0.00	0.00	54.19	118.27	200.10
Water	121.80	164.90	185.00	181.60	192.00	247.00
Superplasticizer	0.00	0.00	6.35	6.23	10.16	32.20
Coarse-Aggregate	801.00	932.00	968.00	972.90	1029.40	1145.00
Fine-Aggregate	594.00	731.00	779.50	773.60	824.00	992.60
Age	1.00	7.00	28.00	45.66	56.00	365.00

Regarding the provided summary, the Min column displays the minimum values, which are 0.0 for Blast Furnace Slag, Fly Ash, and Superplasticizer. The 1st quartile, representing the lower range of values, signifies the point below which 25% of the data lies. As an illustration, for the element "Superplasticizer," the first quartile is "0.00." In this dataset, the median of "Concrete Compressive Strength" is 34.44, while the mean is 35.81. The 3rd quartile, indicating the upper range of values where 75% of the data falls, is exemplified by "Coarse Aggregate," with the third quartile at 1029.40. The Max column shows the maximum value for this category, which is 1145.00.

4.1.1 Correlation Plot

The correlation plot matrix displayed in Figure 2 within the appendix illustrates the relationships among numeric variables. A negative correlation of -0.28 is observed between Concrete compressive strength and Blast furnace slag, indicating that an increase in blast furnace slag is associated with a decrease in Concrete compressive strength. Similarly, there are negative correlations between Concrete compressive strength and fine aggregate (-0.22), coarse aggregate (-0.11), water (-0.08), and Fly ash (-0.40). Furthermore, positive linear correlations are identified between Concrete compressive strength and age (0.08), superplasticizer (0.09), and cement (0.50), signifying that an increase in these variables is linked to an increase in Concrete compressive strength.

4.2 Linear Regression Analysis

In this segment, we present the outcomes of a regression model investigating the correlation between the response variable, Concrete compressive strength, and a set of explanatory variables, including Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age. Table.2 displays the estimated coefficients and corresponding p-values.

Our findings suggest that a one-kilogram increase in cement leads to a $(1.132 \times 10^{-1}) = 0.1132$ increase in Concrete Compressive strength, with other variables held constant. Similarly, each unit increase in Fly Ash corresponds to an increase of $(7.666 \times 10^{-2}) = 0.0766$ in Concrete Compressive strength. The remarkably low p-values for Cement, Blast Furnace Slag, and Superplasticizer signify their significance as predictors for Concrete Compressive strength.

Table 2: Regression Coefficients

Variable	Estimate	P-value
(Intercept)	-3.129×10^{0}	0.89171
Cement	1.132×10^{-1}	$< 2 \times 10^{-16}$
Blast Furnace Slag	9.668×10^{-2}	$< 2 \times 10^{-16}$
Fly Ash	7.666×10^{-2}	3.42×10^{-12}
Water	-1.760×10^{-1}	4.75×10^{-7}
Superplasticizer	2.253×10^{-1}	0.00539
Coarse-Aggregate	8.448×10^{-3}	0.29881
Fine-Aggregate	1.138×10^{-2}	0.21883
‘Age (day)‘	2.517×10^{-1}	$< 2 \times 10^{-16}$
Age Cube	-1.471×10^{-6}	$< 2 \times 10^{-16}$

4.3 Model Selection

In the subsequent steps, the model exhibiting the optimal performance based on AIC and BIC is chosen. Consequently, the approach employed for model selection is the *Best Subset Selection*, which involves evaluating all potential combinations of the explanatory variables. The top-performing model is then selected, and AIC and BIC values are computed for each of the models.

We choose models based on the lowest AIC and BIC values. As indicated in Table 3, it is apparent that *model7* exhibits the lowest AIC value, specifically 7453.97, and also the lowest BIC value, which stands at 4531.08. Hence, *model7*, characterized by seven covariates, is selected as the model with the minimum AIC and BIC values.

Table 3: Result of Best Subset selection.

	Predictors	AIC	BIC
1	Cement	8435.11	5509.68
2	Cement Superplasticizer	8285.04	5359.29
3	Cement Superplasticizer ‘Age (day)‘	8055.64	5130.48
4	Cement Superplasticizer ‘Age (day)‘ AgeCube	7841.66	4915.33
5	Cement Blast-Furnace-Slag Water ‘Age (day)‘ AgeCube	7616.11	4691.36
6	Cement Blast-Furnace-Slag Fly-Ash Water ‘Age (day)‘ AgeCube	7460.14	4537.13
7	Cement Blast-Furnace-Slag Fly-Ash Water Superplasticizer ‘Age (day)‘ AgeCube	7453.97	4531.08
8	Cement Blast-Furnace-Slag Fly-Ash Water Superplasticizer Fine-Aggregate ‘Age (day)‘ AgeCube	7455.53	4532.67
9	Cement Blast-Furnace-Slag Fly-Ash Water Superplasticizer Coarse-Aggregate Fine-Aggregate ‘Age (day)‘ AgeCube	7456.44	4533.62

Now, we analyze the parameter estimates derived from the model with the smallest AIC and BIC. The estimated parameter values, along with their associated p-values and confidence intervals, are presented in Table.4.

Our results indicate that an increase of one kilogram in cement is associated with a 28.79 MPa increase in Concrete Compressive strength when all other variables are set to zero. Likewise, a one-kilogram increase in Blast Furnace slag results in a 0.92 MPa increase in Concrete Compressive strength, assuming other variables remain constant. Conversely, a one-kilogram increase in water leads to a -0.22 MPa decrease in Concrete Compressive strength when other variables are held at zero. Similar interpretations can be extended to the remaining parameter estimates.

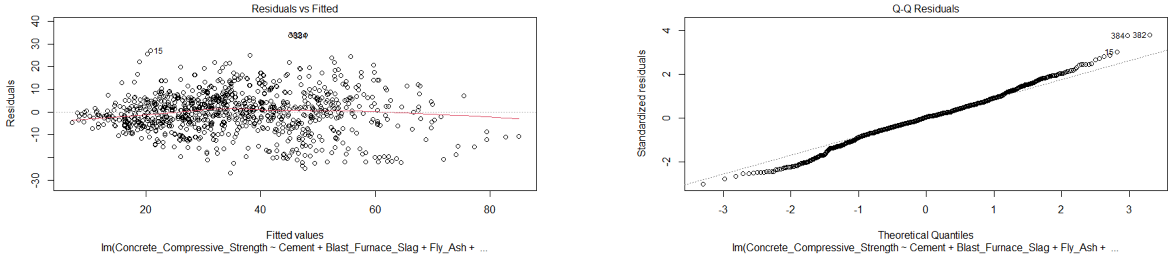
Table 4: Linear regression Parameter estimates, p-values and Confidence Intervals

Variables	Estimates	P-values	Confidence Interval
(Intercept)	2.879×10^{1}	$< 2e - 16$	[22.51, 35.07]
Cement	1.095×10^{-1}	$< 2e - 16$	[0.10, 0.11]
Blast Furnace Flag	9.249×10^{-2}	$< 2e - 16$	[0.08, 0.10]
Fly Ash	6.420672×10^{-2}	$< 2e - 16$	[0.06, 0.08]
Water	-2.709766×10^{-1}	$< 2e - 16$	[-0.27, -0.21]
Age (day)	2.354581×10^{-1}	$< 2e - 16$	[0.23, 0.26]
Age Cube	-1.635630×10^{-6}	$< 2e - 16$	[-0.000001, -0.000001]

Considering the 95% confidence interval, it is evident that none of the parameters include zero within their ranges. The absence of zero within a parameter's interval would imply that the respective variable is statistically significant in influencing Concrete compressive strength.

4.3.1 Heteroscedasticity, Normality and Multicollinearity

In Figure 1(a), the presence of heteroscedasticity is illustrated by the red line in the plot, indicating a non-uniform variance and variability in the data points. The spread of the data suggests that the variability of residuals or errors changes as the predicted values either increase or decrease. In simpler terms, the inconsistency in the spread of residuals across different predicted values is referred to as heteroscedasticity in variance. This implies that the data's variability is not consistent across the entire range of predicted values. Moreover, based on the observation from Figure 1(b), we can infer that the residuals demonstrate a normal distribution, as a majority of points align along a straight line.



(a) Plot 1

(b) Plot 2

Figure 1: Residual and Q-Q plot for normality assumption

Regarding the multicollinearity assumption, it is met as none of the independent variables have a VIF value exceeding 10, indicating the absence of multicollinearity. The results are shown in Table 5.

Table 5: Multicollinearity

Variables	VIF-Value
Cement	1.594228
Blast Furnace Flag	1.444797
Fly Ash	1.750451
Water	1.205932
Age (day)	3.853179
Age Cube	3.819745

The Final model regression can be formulated as:

$$\begin{aligned}\mathbf{y}_{\text{concrete-compressive strength}} &= \beta_0 + x_{\text{Cement}}\beta_1 \\ &\quad + x_{\text{Blast-Furnace-Slag}}\beta_2 + x_{\text{Fly-Ash}}\beta_3 \\ &\quad + x_{\text{Water}}\beta_4 + x_{\text{Age}}\beta_5 \\ &\quad + x_{\text{AgeCube}}\beta_6 + \varepsilon.\end{aligned}$$

Taking into account the issues associated with the model, linear regression relies on the assumption of a linear relationship between predictor variables and the response variable. If this assumption is not met as it shows non-linearity, the model might produce biased or inefficient estimates.

5 Summary

The original information stems from concrete mix measurements conducted in a laboratory at various time points. The dataset under examination comprises 1,030 instances, encompassing nine different factors. The data integrity is high, with no missing values, indicating overall good quality. The analysis aims to determine the most suitable regression model elucidating the relationship between Concrete Compressive Strength and other ingredients such as Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, and Fine Aggregate.

A descriptive analysis revealed an average Concrete Compressive Strength of 34.44. The correlation plot matrix highlighted negative correlations between Concrete Compressive Strength and certain variables such as Blast Furnace Slag, Fine Aggregate, Coarse Aggregate, Water, and Fly Ash. Conversely, positive linear correlations were noted between Concrete Compressive Strength and variables like Age, Superplasticizer, and Cement, suggesting that an increase in these factors corresponds to a rise in Concrete Compressive Strength.

Following the descriptive analysis, a regression model incorporating all variables was developed. Utilizing the Best Subset method with AIC and BIC measures, a model featuring seven variables (Cement, Blast-Furnace-Slag, Fly-Ash, Water, Superplasticizer, Age (day), AgeCube) was determined to offer a superior fit. Coefficients of the estimated parameters were interpreted based on AIC and BIC, and the model's adherence to regression assumptions was assessed through techniques like a Q-Q plot, revealing heteroscedasticity. The assumption of multicollinearity was found to be satisfied, as none of the independent variables exhibited a VIF value exceeding 10, indicating the absence of multicollinearity.

In future research endeavors, the methodologies employed in this project can be extended to encompass a substantially larger dataset, potentially sourced from various cement manufacturers. It would be advantageous to incorporate additional variables such as mixing duration, temperature, and humidity levels to enhance the depth and scope of the analysis.

Bibliography

Thomas Lumley based on Fortran code by Alan Miller. *leaps: Regression Subset Selection*, 2020. URL `https://CRAN.R-project.org/package=leaps`. R package version 3.1.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression: Models, Methods and Application*. Springer, Berlin, Heidelberg, 2013. doi: https://doi.org/10.1007/978-3-642-34333-9.

John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition, 2019. URL `https://socialsciences.mcmaster.ca/jfox/Books/Companion/`.

Aravind Hebbali. *olsrr: Tools for Building OLS Regression Models*, 2020. URL `https://CRAN.R-project.org/package=olsrr`. R package version 0.5.3.

Richard M Heiberger and Burt Holland. *Statistical Analysis and Data Display*. Springer, New York, NY, 2015. doi: https://doi.org/10.1007/978-1-4939-2122-5.

Gareth James. *Introduction to Statistical Learning*. Springer, New York, 1 edition, 2013. ISBN 978-1461471370.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

Taiyun Wei and Viliam Simko. *R package 'corrplot': Visualization of a Correlation Matrix*, 2021. URL `https://github.com/taiyun/corrplot`. (Version 0.92).

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Hadley Wickham and Jennifer Bryan. *readxl: Read Excel Files*, 2023. URL `https://CRAN.R-project.org/package=readxl`. R package version 1.4.3.

Appendix

A Additional figures

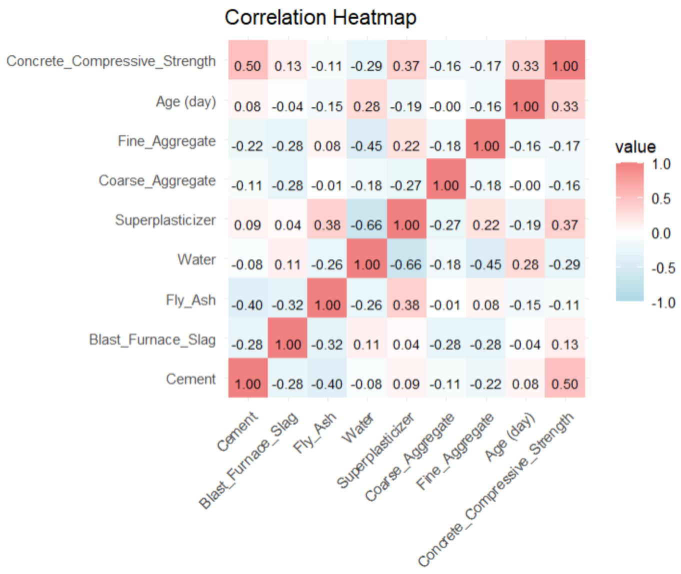


Figure 2: Correlation Plot

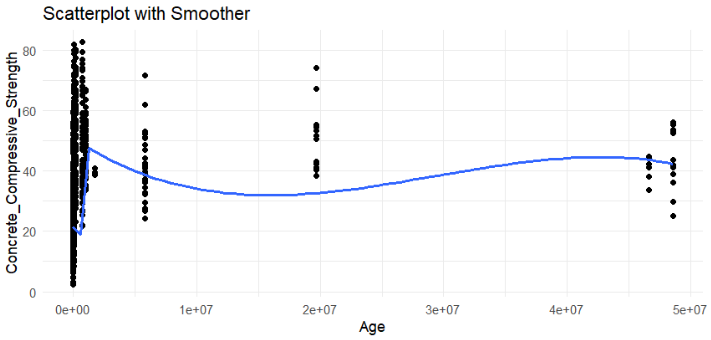


Figure 3: Non-Linearity Age