DATA2002 Project Report *https://github.sydney.edu.au/ejes2694/CC08E6*

Abstract

This investigation was conducted to determine the effects of component concentration on the compressive strength of concrete. A simple linear, full, backwards and forwards regression model was built. Model assumptions were checked to increase the validity of the study with the final predicted equation for concrete strength to be:

$$\widehat{\text{strength}} = -23.16 + 0.12(\text{cement}) + 0.1(\text{slag}) + 0.09(\text{ash}) - 0.15(\text{water}) + 0.29(\text{superplastic}) + 0.11(\text{age}) + 0.02(\text{coarseagg}) + 0.02(\text{fineagg})$$

The final model forecasts cement, slag, ash and superplastic to positively affect concrete strength, water to have an inverse effect and coarse/fine aggregate to be uninfluential. In future, this study could be improved by employing non-linear regression analysis and adding constraints for necessary conditions for concrete formation such as a positive water value.

Introduction

In this study, we looked at the effect of different concrete components on its compressive strength. We aimed to find which components had the largest influence on concrete strength and in which ways by constructing various linear models based on the data given by UC Irvine. The results of our study could be utilised to calculate the strength of a concrete mixture or to achieve the most optimal formula that maximises concrete compressive strength. The importance of this study is predicated on concrete's paramount role in the infrastructure industry as specific requirements of projects require different strengths of concrete.

Data Set - Description

Overall, this multivariate dataset consisted of 1030 instances with 9 total attributes. The dataset included 8 independent input variables and one dependent output variable with the method of data collecting unknown (UCI, 1998). The dependent variable was concrete compressive strength (MPa)

Independent Variable Units	
Cement	kg/m3
Blast Furnace Slag	kg/m3
Fly Ash	kg/m3
Water	kg/m3
Superplasticizer	kg/m3
Coarse Aggregate	kg/m3
Fine Aggregate	kg/m3
Age	Days

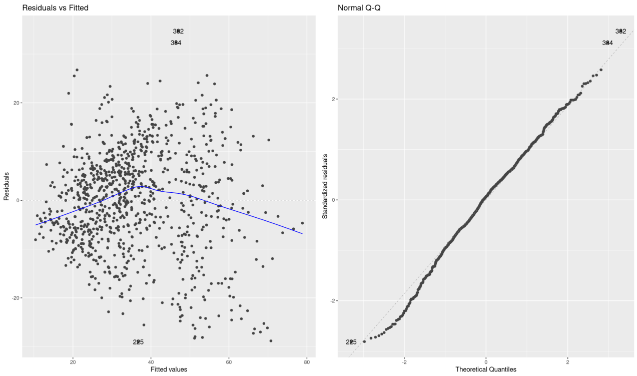
Analysis - EDA

From the boxplot and the histogram (see appendix), it could be seen that the output variable, which is the concrete compressive strength, follows a normal distribution. This variable ranges from 2.33 to 82.6, has a mean of 35.82 and a median of 34.44 MPa. The correlation plot shows that concrete strength has the strongest positive correlation with cement and superplasticizer. The scatter plot indicates that no input variables have visibly strong relationships with the output variable. However, component 1, which is cement, has a noticeable positive trendline, but the points still distribute fairly randomly.

Model Selection

In total, we fitted 4 potential models to the dataset. First, our EDA indicated that cement had the strongest correlation of any of the variables with concrete strength. Accordingly, we started with a simple linear regression model that used cement as the only predictor. Following this, a full linear model was created that took into account all 8 variables. Finally, we sought to generate some models using feature selection, so forwards and backwards feature selection models were fitted.

Assumption Checking



Residuals vs fitted plot and normal qq plot were used to check the model against the 4 assumptions: independence, linearity, homoscedasticity, and normality.

1. **Independence:** There exists independence between the errors. This was covered during the experimental design stage before the data collection.
2. **Linearity**: From the residuals vs fitted plot, the residuals appear to be equally distributed above and below zero. This shows that there are no clear patterns and the model used was not misspecified.
3. **Homoscedasticity:** From the residuals vs fitted plot, the residuals do not seem to be spread out or change the variability over the range of the fitted values, which means that the constant error variance assumption is met.
4. **Normality:** From the normal qq plot, it depicts that points are reasonably close to the diagonal. Although the top and bottom points may slightly depart from the diagonal, this can be relaxed with the CLM since the sample size is sufficiently large enough.

Results & Discussion

	Root Mean Square of Errors	Mean Absolute Value of Errors
Cement Model	15.26	12.45
Full Model	11.49	9.04
Forward Model	10.37	8.26
Backward Model	10.35	8.21

The evaluation of models was completed with a 10-fold cross validation, with observations split into 10 randomly allocated folds. Thus, no stratification was used.

As an aside, we looked at whether feature selection stayed consistent over training on the different folds. We investigated this using a forward feature selection technique. 8 out of 10 times, model construction chose cement, slag, ash, water, superplastic and age; leaving out coarse and fine aggregate. This implies that it is possible that coarse and fine aggregate have a negligible effect on the strength of concrete whereas the other 6 variables have more influence.

It was found that the model that used backwards feature selection was the most accurate in prediction. This technique chose to include all 8 variables in the model. This model held the smallest root mean square of errors with 10.35 as well as the smallest mean absolute value of errors with 8.21. This was the equation of the model produced.

A t-test was used to check if the coefficients were significantly different to 0 (and thus there was a relationship between the given variable and the strength). All p-values except for coarse and fine aggregate were well below 0.05, and thus significant. Both coarse and fine aggregate were just above 0.05, which seems to be consistent with the way our various models tended to make a borderline decision on whether they were worth including (e.g. the forward model removed them, the backwards model did not). Thus we can reject the null hypotheses for all coefficients (except fine and coarse aggregate) that are equal to 0.

	Concrete Strength (MPa)	
Predictors	*Estimates*	*p*
(Intercept)	-23.16	0.384
cement	0.12	**<0.001**
slag	0.10	**<0.001**
ash	0.09	**<0.001**
water	-0.15	**<0.001**
superplastic	0.29	**0.002**
age	0.11	**<0.001**
coarseagg	0.02	0.055
fineagg	0.02	0.060
Observations	1030	
R^2 / R^2 adjusted	0.615 / 0.612	

Looking at this model, it can be seen that all variables except for water had a positive coefficient, meaning that decreasing the amount of water would increase the strength, but increasing any of the other variables would have a positive impact on the strength. Interestingly, the superplastic has the highest coefficient of all of the variables (holding all other variables constant, a unit increase in superplastic would increase the concrete strength by approximately 0.29 megapascals), meaning that a unit increase in superplastic would have more of an impact in the strength of the concrete than any of the other variables. Fine and coarse aggregate have extremely small coefficients of 0.02, which does seem to be consistent with the fact that they were sometimes ignored when feature selection was undertaken. This, along with the EDA and assumption checking leads us to say that there is strong evidence to conclude a linear relationship between the 7 concrete components + time, and the strength of the concrete

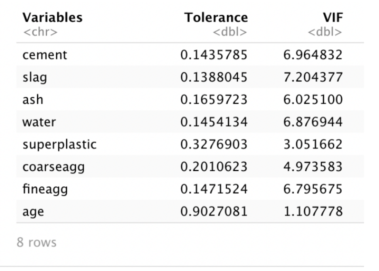
The forwards model was second, with an error rate that was approximately 0.02. The full model, which had a significantly higher error rate, was next. Finally, the simple model using cement as the single predictor had the highest error rate.

Collinearity

To ensure the validity of our model, we had to check if there existed collinearity between any of our variables. Through collinearity diagnosis between two or more variables from previously fitted linear regression models, we seek to determine whether collinearity exists. Variance Inflation Factors (VIF) will be examined to check whether collinearity exists.

As mentioned in a previous study, if the Tolerance is low, the other independent variables should explain an increase or decrease in the value of the independent variable of interest. Hence, the variables are correlated and therefore multicollinearity may exist." (Coding prof, 2022). We will be using this as a comparable factor to briefly determine whether collinearity exists in our study.

Package olsrr will be used to check VIF and tolerance level, and "strength to all " multiple regression will be tested for colinearity as collinearity diagnosis requires two or more variables.



Variables <chr>	Tolerance <dbl>	VIF <dbl>
cement	0.1435785	6.964832
slag	0.1388045	7.204377
ash	0.1659723	6.025100
water	0.1454134	6.876944
superplastic	0.3276903	3.051662
coarseagg	0.2010623	4.973583
fineagg	0.1471524	6.795675
age	0.9027081	1.107778

8 rows

From the returned stat, we can tell that collinearity does not exist on age as it has a VIF value of 1 which means that the regression coefficient is not inflated by the presence of the other predictors, hence indicating collinearity does not exist.

It is apparent in the collinearity diagnosis, that cement, slag, ash and water showed visible contributing correlation with the compressive strength, as they both have a high VIF value that are greater than 4. However, if only determined by the VIF values, collinearity is detected in every single variable in the given dataset apart from age and superplastic. With varying standards on how VIF should be interpreted correctly, further investigation is required to determine their impacts.

Conclusion:

In conclusion, it seems that all 8 variables have a linear effect on concrete strength, with superplasticizer being most influential and fine & coarse aggregate being the least influential. Our chosen model predicts concrete strength on average within 8-10 MPa of its real strength. Our model has some limitations however

First, our linear model has a negative intercept, which describes that if no components were put in the concrete, it would have a strength of -23.16. This is not ideal in two respects. Both that a negative strength is not possible, and if no components are put in the concrete, then you don't have concrete. Thus, a linear model is not able to represent this relationship. Perhaps this could be fixed by forcing the intercept to be at the origin, or ensuring that the model is never used in this unrealistic range.

Second, there are likely many other factors that affect the strength of concrete. These include the weather conditions where it is mixed and poured (humidity and heat both affect drying times and strength), the quality of components used, and the way in which the concrete is mixed and poured. Perhaps to create a more accurate model, these features should be included in the dataset.

Finally, our model doesn't take into consideration the interactions of different components in concrete. For instance, our model implies that adding water reduces concrete strength, and so it would follow that adding no water should increase the strength. However, if no water is mixed into the concrete, then the mix is just powder, and isn't concrete at all. Perhaps some checking could be done before an instance is predicted to ensure that it meets the essential component requirements to be concrete.

Bibliography

Anderson, Daniel, Andrew Heiss, and Jay Sumners. 2022. Equatiomatic: Transform Models into LaTeX Equations. https://CRAN.R-project.org/package=equatiomatic.

Collinearity Diagnostics, Model Fit & Variable Contribution. (n.d.). Retrieved November 3, 2022, from https://cran.r-project.org/web/packages/olsrr/vignettes/regression_diagnostics.html

Hebbali, Aravind. 2020. Olsrr: Tools for Building OLS Regression Models. https://CRAN.R-project.org/package=olsrr.

Importance of Setting Seed in Model Fitting. (2018, October 1). Kaggle. https://www.kaggle.com/code/obrienmitch94/importance-of-setting-seed-in-model-fitting

Kuhn, Max. 2022. Caret: Classification and Regression Training. https://github.com/topepo/caret/.

Lumley, Thomas. 2020. Leaps: Regression Subset Selection. https://CRAN.R-project.org/package=leaps.

Müller, Kirill, and Hadley Wickham. 2022. Tibble: Simple Data Frames. https://CRAN.R-project.org/package=tibble.

Peterson, Brian G., and Peter Carl. 2020. PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis. https://github.com/braverock/PerformanceAnalytics.

Prof, C. (2022, May 16). 3 ways to test for multicollinearity in R. Retrieved November 2, 2022, from https://www.codingprof.com/3-ways-to-test-for-multicollinearity-in-r-examples/

R Core Team. 2022. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rich, Benjamin. 2021. Table1: Tables of Descriptive Statistics in HTML. https://github.com/benjaminrich/table1.

Sarkar, Deepayan. 2008. Lattice: Multivariate Data Visualization with r. New York: Springer. http://lmdvr.r-forge.r-project.org.

———. 2021. Lattice: Trellis Graphics for r. http://lattice.r-forge.r-project.org/.

UCI. (1998). Concrete Compressive Strength Data Set. *Machine Learning Repository.* Available at: https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength

Wei, Taiyun, and Viliam Simko. 2021a. Corrplot: Visualization of a Correlation Matrix. https://github.com/taiyun/corrplot.

———. 2021b. R Package 'Corrplot': Visualization of a Correlation Matrix. https://github.com/taiyun/corrplot.

Wickham, Hadley. 2016. Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org. 2022a. Forcats: Tools for Working with Categorical Variables (Factors). https://CRAN.R-project.org/package=forcats.

———. 2022c. Tidyverse: Easily Install and Load the Tidyverse. https://CRAN.R-project.org/package=tidyverse.

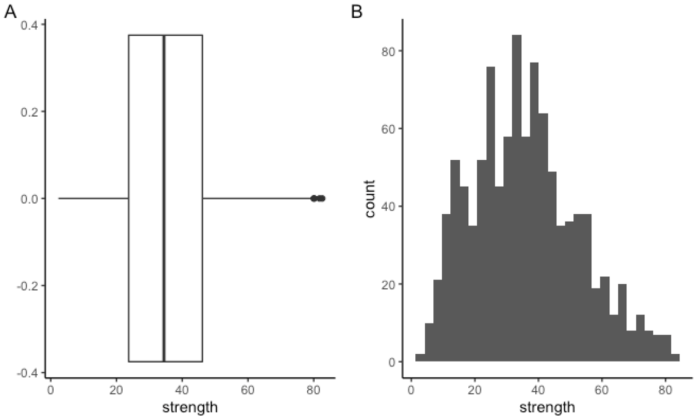
Wickham, et al. 2019. "Welcome to the tidyverse." Journal of Open Source Software 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2022. Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. https://CRAN.R-project.org/package=ggplot2.

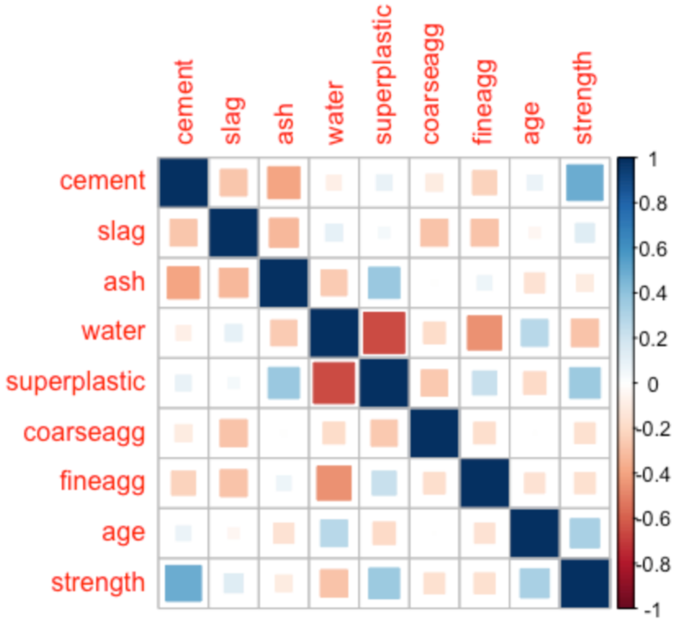
Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. Dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. Readr: Read Rectangular Text Data. https://CRAN.R-project.org/package=readr.

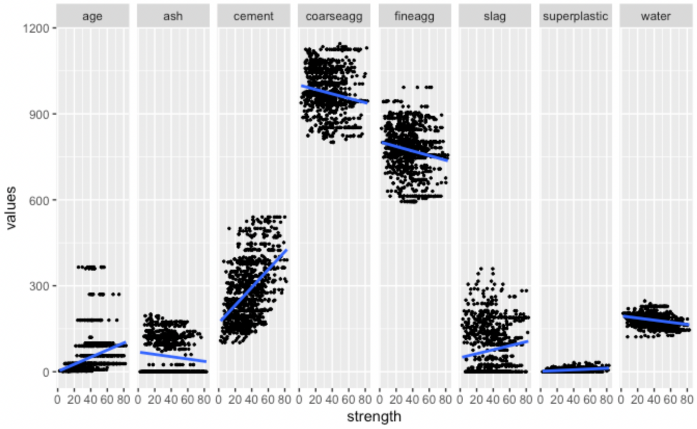
Appendix



Boxplot and Histogram of the output variable - concrete compressive strength



Correlation plot of all variables



Scatter plot of all variables