INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

AQ049-3-M-MMDA

MULTIVARIATE METHODS FOR DATA ANALYSIS

APDMF2112DSBA(DE)(PR)

JULY 2022

TITLE: MULTIPLE LINEAR REGRESSION & FACTOR ANALYSIS

LEE KEAN LIM

TP065778

LECTURER: MR. LOW KOK SUN

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN..... Artificial Neural Network

ANOVA..... Analysis of Variance

CO_2 ..... Carbon Dioxide

MSA..... Measures of Sampling Adequacy

PC ..... Portland Cement

R^2..... Coefficient of Determination

VIF..... Variance Inflation Factor

PART A

MULTIPLE LINEAR REGRESSION

SECTION 1

INTRODUCTION

1.1 INTRODUCTION

Concrete is a construction material which is also the most widely used man made material. It is commonly used in the construction of a structure to provide structural strength, stability, and durability. The fundamental components of concrete are water, aggregates, and Portland cement (PC). However, there have been many innovations in the field of concrete technology to substitute the fundamental components of concrete to materials that are more environmentally sustainable and to achieve better concrete properties. This is due to the manufacturing and utilization of concrete has a significant negative impact to the environment where it contributes to 7% of the total global carbon dioxide (CO_2) emission, which makes this industry one of the biggest CO_2 producers (Watts, 2019). Researchers have utilized various recycled materials such as ground granulated blast furnace slag, fly ash, and granite powder in replacement for PC to produce a concrete that is much more environmentally friendly while maintaining the required concrete properties (Song et al., 2021).

Among the properties of concrete, the compressive strength is the most important property, as it directly influences the safety and durability of the concrete. The compressive strength refers to the load bearing capacity of the concrete before failure. Generally, concrete arrives in batches, where the value of compressive strength for each batch of concrete would varies due to the different material mix ratio that makes up the concrete. The determination of concrete compressive strength is crucial during the construction phase to ensure the concrete is supplying the designed load bearing capacity to ensure the structure is sound safe. The industry practice is to obtain a small sample from every batch of concrete used. Where, these samples are sent to the laboratory for testing to identify the compressive strength after 28 days of curing (Young et al., 2019). This process is highly inefficient and uneconomical as it labor intensive and the test results would only be available after a few weeks. This significantly hinders the construction progress due to the need of waiting for laboratory test results to verify that the concrete has adequate bearing capacity.

The importance of the concrete compressive strength metric has led to the attention of researchers to develop prediction models to predict the compressive strength value. Particularly,

the regression analysis which is one of the statistical methods to produce a prediction has been widely utilized. The regression analysis examines the relationship between one or more independent variables to a single dependent variable. Jin et al. (2018) compared two statistical models based on linear and non-linear regression analysis to predict the concrete compressive strength. It was identified that the non-linear model performs better than the linear model in predicting the concrete compressive strength. In addition, it was mentioned that the non-linear model achieved a relatively high coefficient of determination (R^2) of 0.934 while the linear model achieved only 0.907. Several advantages of using statistical model to predict the concrete compressive strength was mentioned, which include easy and quick model creation, reduction in reliance on laboratory testing, and high reliability in prediction results. Based on the study, it can be identified that the prediction outcome of the concrete compressive strength is affected by a multitude of factors and is non-linear in nature. However, in the works of Chithra et al. (2016), a contradictory results were obtained. The author compared the performance of utilizing multiple regression models and Artificial Neural Network (ANN) models to predict the concrete compressive strength. It was identified that the multiple regression models obtained on average a R^2 of 0.659 while the ANN models obtained on average a R^2 greater than 0.9. The author mentioned that the low R^2 in the regression models can be attributed due to multi-collinearity.

This study investigates the use of multiple linear regression to develop a prediction model to predict the concrete compressive strength based on several concrete constituents provided. The regression model would undergo a stepwise method to identify the significant independent variables for the final model.

1.2 PROBLEM STATEMENT

Based on industry practice, to obtain the test results from a concrete compressive strength test, it requires the wait of 28 days to ensure the concrete curing process has been achieved to reflect the 99% strength of the concrete. The long waiting time for the test results hinders the construction progress due to the safety requirements to verify the concrete strength before the next layer of concrete pouring to ensure the adequacy of the concrete to withstand the structure (Young et al., 2019). The concrete testing phase is both labor intensive and time consuming thus a better solution is to be devised which is the use of a prediction model to predict the concrete compressive strength based on the constituents of the concrete mix. The use of

statistical model such as multiple linear regression can be utilized to predict the concrete compressive strength, which Jin et al. (2018) and Chithra et al. (2016) has utilized such models to predict the compressive strength and achieved reliable results. The use of statistical models to predict the concrete compressive strength would reduce the reliance on labors and reduce the costs for conducting the tests. In addition, statistical models can be easily and quickly developed to predict the concrete compressive strength while providing an understanding of the relationship between the predictors and the output.

1.3 RESEARCH OBJECTIVES

Following outline three research objectives and their corresponding hypothesis to be examined in this study. In overall, the objectives are to identify how each ingredient used in mixing concrete affects the concrete compressive strength.

1. To identify the relationship between water mix quantity and the concrete compressive strength.
   H_o: Water does not have an effect on the concrete compressive strength.
   H_1: Water has an effect on the concrete compressive strength.

2. To identify the relationship between blast furnace slag mix quantity and the concrete compressive strength.
   H_o: Blast furnace slag does not have an effect on the concrete compressive strength.
   H_1: Blast furnace slag has an effect on the concrete compressive strength.

3. To identify the relationship between fly ash mix quantity and the concrete compressive strength.
   H_o: Fly ash does not have an effect on the concrete compressive strength.
   H_1: Fly ash has an effect on the concrete compressive strength.

4. To identify the relationship between coarse aggregate mix quantity and the concrete compressive strength.
   H_o: Coarse aggregate does not have an effect on the concrete compressive strength.
   H_1: Coarse aggregate has an effect on the concrete compressive strength.

SECTION 2

MULTIPLE LINEAR REGRESSION

2.1 DATASET

The dataset used in this study is a secondary dataset from Yeh (1998) which consist of 1030 observations of concrete compressive strength based on different ingredient mix of the concrete. There are nine features in the dataset where all features are metric. Table 2.1 shows the metadata of the dataset which describes the feature label, unit of measurement, and variable type. It is identified that the concrete compressive strength is the dependent variable in this study. Figure 2.1 shows the variable view of the dataset consisting of the metadata from SPSS, while Figure 2.2 shows the snapshot of the dataset in the data view from SPSS.

Table 2.1: Dataset metadata

Feature Name	Label	Unit of Measurement	Variable Type
X1	Cement	Kg/m^3	Independent
X2	Blast Furnace Slag	Kg/m^3	Independent
X3	Fly Ash	Kg/m^3	Independent
X4	Water	Kg/m^3	Independent
X5	Superplasticizer	Kg/m^3	Independent
X6	Coarse Aggregate	Kg/m^3	Independent
X7	Fine Aggregate	Kg/m^3	Independent
X8	Age	Day	Independent
Y	Concrete Compressive Strength	MPa	Dependent

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	X1	Numeric	5	1	Cement	None	None	8	Right	Scale	Input
2	X2	Numeric	5	1	Blast Furnance Slag	None	None	8	Right	Scale	Input
3	X3	Numeric	5	1	Fly Ash	None	None	8	Right	Scale	Input
4	X4	Numeric	5	1	Water	None	None	8	Right	Scale	Input
5	X5	Numeric	4	1	Superplasticizer	None	None	8	Right	Scale	Input
6	X6	Numeric	6	1	Coarse Aggregate	None	None	8	Right	Scale	Input
7	X7	Numeric	5	1	Fine Aggregate	None	None	8	Right	Scale	Input
8	X8	Numeric	3	0	Age	None	None	8	Right	Scale	Input
9	Y	Numeric	5	2	Strength	None	None	8	Right	Scale	Input

Figure 2.1: Variable view from SPSS

	X1	X2	X3	X4	X5	X6	X7	X8	Y
472	446.0	24.0	79.0	162.0	11.6	967.0	712.0	28	57.03
473	446.0	24.0	79.0	162.0	11.6	967.0	712.0	28	44.42
474	446.0	24.0	79.0	162.0	11.6	967.0	712.0	28	51.02
475	446.0	24.0	79.0	162.0	10.3	967.0	712.0	28	53.39
476	446.0	24.0	79.0	162.0	11.6	967.0	712.0	3	35.36
477	446.0	24.0	79.0	162.0	11.6	967.0	712.0	3	25.02
478	446.0	24.0	79.0	162.0	11.6	967.0	712.0	3	23.35
479	446.0	24.0	79.0	162.0	11.6	967.0	712.0	7	52.01
480	446.0	24.0	79.0	162.0	11.6	967.0	712.0	7	38.02
481	446.0	24.0	79.0	162.0	11.6	967.0	712.0	7	39.30
482	446.0	24.0	79.0	162.0	11.6	967.0	712.0	56	61.07
483	446.0	24.0	79.0	162.0	11.6	967.0	712.0	56	56.14
484	446.0	24.0	79.0	162.0	11.6	967.0	712.0	56	55.25
485	446.0	24.0	79.0	162.0	10.3	967.0	712.0	56	54.77
486	387.0	20.0	94.0	157.0	14.3	938.0	845.0	28	50.24
487	387.0	20.0	94.0	157.0	13.9	938.0	845.0	28	46.68
488	387.0	20.0	94.0	157.0	11.6	938.0	845.0	28	46.68
489	387.0	20.0	94.0	157.0	14.3	938.0	845.0	3	22.75
490	387.0	20.0	94.0	157.0	13.9	938.0	845.0	3	25.51
491	387.0	20.0	94.0	157.0	11.6	938.0	845.0	3	34.77
492	387.0	20.0	94.0	157.0	14.3	938.0	845.0	7	36.84
493	387.0	20.0	94.0	157.0	13.9	938.0	845.0	7	45.90
494	387.0	20.0	94.0	157.0	11.6	938.0	845.0	7	41.67
495	387.0	20.0	94.0	157.0	14.3	938.0	845.0	56	56.34
496	387.0	20.0	94.0	157.0	13.9	938.0	845.0	56	47.97
497	387.0	20.0	94.0	157.0	11.6	938.0	845.0	56	61.46
498	355.0	19.0	97.0	145.0	13.1	967.0	871.0	28	44.03
499	355.0	19.0	97.0	145.0	12.3	967.0	871.0	28	55.45
500	491.0	26.0	123.0	210.0	3.9	882.0	699.0	28	55.55
501	491.0	26.0	123.0	201.0	3.9	822.0	699.0	28	57.92
502	491.0	26.0	123.0	210.0	3.9	882.0	699.0	3	25.61
503	491.0	26.0	123.0	210.0	3.9	882.0	699.0	7	33.49
504	491.0	26.0	123.0	210.0	3.9	882.0	699.0	56	59.59
505	491.0	26.0	123.0	201.0	3.9	822.0	699.0	3	29.55
506	491.0	26.0	123.0	201.0	3.9	822.0	699.0	7	37.92
507	491.0	26.0	123.0	201.0	3.9	822.0	699.0	56	61.86
508	424.0	22.0	132.0	178.0	8.5	822.0	750.0	28	62.05
509	424.0	22.0	132.0	178.0	8.5	882.0	750.0	3	32.01
510	424.0	22.0	132.0	168.0	8.9	822.0	750.0	28	72.10
511	424.0	22.0	132.0	178.0	8.5	822.0	750.0	7	39.00
512	424.0	22.0	132.0	178.0	8.5	822.0	750.0	56	65.70
513	424.0	22.0	132.0	168.0	8.9	822.0	750.0	3	32.11
514	424.0	22.0	132.0	168.0	8.9	822.0	750.0	7	40.29
515	424.0	22.0	132.0	168.0	8.9	822.0	750.0	56	74.36
516	202.0	11.0	141.0	206.0	1.7	942.0	801.0	28	21.97
517	202.0	11.0	141.0	206.0	1.7	942.0	801.0	3	9.85
518	202.0	11.0	141.0	206.0	1.7	942.0	801.0	7	15.07
519	202.0	11.0	141.0	206.0	1.7	942.0	801.0	56	23.25
520	284.0	15.0	141.0	179.0	5.5	842.0	801.0	28	43.73
521	284.0	15.0	141.0	179.0	5.5	842.0	801.0	3	13.40
522	284.0	15.0	141.0	179.0	5.5	842.0	801.0	7	24.13

Figure 2.2: Data view from SPSS

2.2 RESULT AND ANALYSIS

2.2.1 Stepwise Linear Multiple Regression

This study utilizes the stepwise multiple linear regression method to identify the significant independent variables to be included into the model. In addition, this study adopts the assumption of 95% significance level for all testing. The following outline the stepwise linear regression model development.

Step 0: Base model

To identify the first variable to be included into the model, the correlation coefficient is used. Table 2.2 shows the correlation matrix among all variables from the dataset.

Table 2.2: Full correlation matrix

Correlations

		Cement	Blast Furnance Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Age	Strength
Cement	Pearson Correlation	1	-.275**	-.397**	-.082**	.092**	-.109**	-.223**	.082**	.498**
	Sig. (2-tailed)		.000	.000	.009	.003	.000	.000	.009	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
Blast Furnance Slag	Pearson Correlation	-.275**	1	-.324**	.107**	.043	-.284**	-.282**	-.044	.135**
	Sig. (2-tailed)	.000		.000	.001	.165	.000	.000	.156	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
Fly Ash	Pearson Correlation	-.397**	-.324**	1	-.257**	.378**	-.010	.079*	-.154**	-.106**
	Sig. (2-tailed)	.000	.000		.000	.000	.750	.011	.000	.001
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
Water	Pearson Correlation	-.082**	.107**	-.257**	1	-.658**	-.182**	-.451**	.278**	-.290**
	Sig. (2-tailed)	.009	.001	.000		.000	.000	.000	.000	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
Superplasticizer	Pearson Correlation	.092**	.043	.378**	-.658**	1	-.266**	.223**	-.193**	.366**
	Sig. (2-tailed)	.003	.165	.000	.000		.000	.000	.000	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
Coarse Aggregate	Pearson Correlation	-.109**	-.284**	-.010	-.182**	-.266**	1	-.178**	-.003	-.165**
	Sig. (2-tailed)	.000	.000	.750	.000	.000		.000	.923	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
Fine Aggregate	Pearson Correlation	-.223**	-.282**	.079*	-.451**	.223**	-.178**	1	-.156**	-.167**
	Sig. (2-tailed)	.000	.000	.011	.000	.000	.000		.000	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
Age	Pearson Correlation	.082**	-.044	-.154**	.278**	-.193**	-.003	-.156**	1	.329**
	Sig. (2-tailed)	.009	.156	.000	.000	.000	.923	.000		.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
Strength	Pearson Correlation	.498**	.135**	-.106**	-.290**	.366**	-.165**	-.167**	.329**	1
	Sig. (2-tailed)	.000	.000	.001	.000	.000	.000	.000	.000	
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

The "Strength" variable is the dependent variable, while all other variables are the independent variables. Based on the correlation matrix, the "Cement" variable exhibits the highest correlation to the "Strength" with a correlation magnitude of 0.498 as compared to other independent variables. In addition, having a significance less than 0.05. Therefore, the "Cement" variable will be the first variable to enter the model.

Step 1: Model 1

Variable added from previous step: “Cement”

To identify the next significant independent variable, the partial correlation is used. Table 2.3 shows the partial correlation matrix after excluding the variable added from previous step.

Table 2.3: Partial correlation matrix for model 1

Correlations

Control Variables			Blast Furnance Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Age	Strength
Cement	Blast Furnance Slag	Correlation	1.000	-.491	.088	.072	-.329	-.366	-.023	.326
		Significance (2-tailed)	.	.000	.004	.021	.000	.000	.468	.000
		df	0	1027	1027	1027	1027	1027	1027	1027
	Fly Ash	Correlation	-.491	1.000	-.316	.453	-.059	-.011	-.133	.116
		Significance (2-tailed)	.000	.	.000	.000	.060	.736	.000	.000
		df	1027	0	1027	1027	1027	1027	1027	1027
	Water	Correlation	.088	-.316	1.000	-.655	-.193	-.483	.286	-.288
		Significance (2-tailed)	.004	.000	.	.000	.000	.000	.000	.000
		df	1027	1027	0	1027	1027	1027	1027	1027
	Superplasticizer	Correlation	.072	.453	-.655	1.000	-.259	.251	-.202	.371
		Significance (2-tailed)	.021	.000	.000	.	.000	.000	.000	.000
		df	1027	1027	1027	0	1027	1027	1027	1027
	Coarse Aggregate	Correlation	-.329	-.059	-.193	-.259	1.000	-.209	.006	-.128
		Significance (2-tailed)	.000	.060	.000	.000	.	.000	.848	.000
		df	1027	1027	1027	1027	0	1027	1027	1027
	Fine Aggregate	Correlation	-.366	-.011	-.483	.251	-.209	1.000	-.142	-.067
		Significance (2-tailed)	.000	.736	.000	.000	.000	.	.000	.032
		df	1027	1027	1027	1027	1027	0	1027	1027
	Age	Correlation	-.023	-.133	.286	-.202	.006	-.142	1.000	.333
		Significance (2-tailed)	.468	.000	.000	.000	.848	.000	.	.000
		df	1027	1027	1027	1027	1027	1027	0	1027
	Strength	Correlation	.326	.116	-.288	.371	-.128	-.067	.333	1.000
		Significance (2-tailed)	.000	.000	.000	.000	.000	.032	.000	.
		df	1027	1027	1027	1027	1027	1027	1027	0

Based on the partial correlation matrix, the “Superplasticizer” variable exhibits the highest correlation to the “Strength” with a partial correlation magnitude of 0.371 as compared to other independent variables. In addition, having a significance less than 0.05. Therefore, the “Superplasticizer” variable will be the next variable to enter the model.

Step 2: Model 2

Variables added from previous steps: “Cement”, “Superplasticizer”

To identify the next significant independent variable, the partial correlation is used. Table 2.4 shows the partial correlation matrix after excluding the variable added from previous step.

Table 2.4: Partial correlation matrix for model 2

Correlations

Control Variables			Blast Furnance Slag	Fly Ash	Water	Coarse Aggregate	Fine Aggregate	Age	Strength
Cement & Superplasticizer	Blast Furnance Slag	Correlation	1.000	-.589	.180	-.322	-.398	-.008	.323
		Significance (2-tailed)	.	.000	.000	.000	.000	.789	.000
		df	0	1026	1026	1026	1026	1026	1026
	Fly Ash	Correlation	-.589	1.000	-.029	.068	-.144	-.048	-.063
		Significance (2-tailed)	.000	.	.353	.029	.000	.126	.043
		df	1026	0	1026	1026	1026	1026	1026
	Water	Correlation	.180	-.029	1.000	-.496	-.435	.208	-.065
		Significance (2-tailed)	.000	.353	.	.000	.000	.000	.038
		df	1026	1026	0	1026	1026	1026	1026
	Coarse Aggregate	Correlation	-.322	.068	-.496	1.000	-.155	-.049	-.036
		Significance (2-tailed)	.000	.029	.000	.	.000	.118	.248
		df	1026	1026	1026	0	1026	1026	1026
	Fine Aggregate	Correlation	-.398	-.144	-.435	-.155	1.000	-.096	-.177
		Significance (2-tailed)	.000	.000	.000	.000	.	.002	.000
		df	1026	1026	1026	1026	0	1026	1026
	Age	Correlation	-.008	-.048	.208	-.049	-.096	1.000	.449
		Significance (2-tailed)	.789	.126	.000	.118	.002	.	.000
		df	1026	1026	1026	1026	1026	0	1026
	Strength	Correlation	.323	-.063	-.065	-.036	-.177	.449	1.000
		Significance (2-tailed)	.000	.043	.038	.248	.000	.000	.
		df	1026	1026	1026	1026	1026	1026	0

Based on the partial correlation matrix, the “Age” variable exhibits the highest correlation to the “Strength” with a partial correlation magnitude of 0.449 as compared to other independent variables. In addition, having a significance less than 0.05. Therefore, the “Age” variable will be the next variable to enter the model.

Step 3: Model 3

Variables added from previous steps: “Cement”, “Superplasticizer”, “Age”

To identify the next significant independent variable, the partial correlation is used. Table 2.5 shows the partial correlation matrix after excluding the variable added from previous step.

Table 2.5: Partial correlation matrix for model 3

Correlations

Control Variables			Blast Furnance Slag	Fly Ash	Water	Coarse Aggregate	Fine Aggregate	Strength
Cement & Superplasticizer & Age	Blast Furnance Slag	Correlation	1.000	-.590	.186	-.323	-.400	.366
		Significance (2-tailed)	.	.000	.000	.000	.000	.000
		df	0	1025	1025	1025	1025	1025
	Fly Ash	Correlation	-.590	1.000	-.019	.066	-.149	-.047
		Significance (2-tailed)	.000	.	.533	.035	.000	.134
		df	1025	0	1025	1025	1025	1025
	Water	Correlation	.186	-.019	1.000	-.498	-.426	-.181
		Significance (2-tailed)	.000	.533	.	.000	.000	.000
		df	1025	1025	0	1025	1025	1025
	Coarse Aggregate	Correlation	-.323	.066	-.498	1.000	-.160	-.016
		Significance (2-tailed)	.000	.035	.000	.	.000	.612
		df	1025	1025	1025	0	1025	1025
	Fine Aggregate	Correlation	-.400	-.149	-.426	-.160	1.000	-.151
		Significance (2-tailed)	.000	.000	.000	.000	.	.000
		df	1025	1025	1025	1025	0	1025
	Strength	Correlation	.366	-.047	-.181	-.016	-.151	1.000
		Significance (2-tailed)	.000	.134	.000	.612	.000	.
		df	1025	1025	1025	1025	1025	0

Based on the partial correlation matrix, the “Blast Furnace Slag” variable exhibits the highest correlation to the “Strength” with a partial correlation magnitude of 0.366 as compared to other independent variables. In addition, having a significance less than 0.05. Therefore, the “Blast Furnace Slag” variable will be the next variable to enter the model.

Step 4: Model 4

Variables added from previous steps: “Cement”, “Superplasticizer”, “Age”, “Blast Furnace Slag”

To identify the next significant independent variable, the partial correlation is used. Table 2.6 shows the partial correlation matrix after excluding the variable added from previous step.

Table 2.6: Partial correlation matrix for model 4

Correlations

Control Variables			Fly Ash	Water	Coarse Aggregate	Fine Aggregate	Strength
Cement & Superplasticizer & Age & Blast Furnance Slag	Fly Ash	Correlation	1.000	.113	-.163	-.521	.225
		Significance (2-tailed)	.	.000	.000	.000	.000
		df	0	1024	1024	1024	1024
	Water	Correlation	.113	1.000	-.471	-.391	-.272
		Significance (2-tailed)	.000	.	.000	.000	.000
		df	1024	0	1024	1024	1024
	Coarse Aggregate	Correlation	-.163	-.471	1.000	-.334	.116
		Significance (2-tailed)	.000	.000	.	.000	.000
		df	1024	1024	0	1024	1024
	Fine Aggregate	Correlation	-.521	-.391	-.334	1.000	-.005
		Significance (2-tailed)	.000	.000	.000	.	.866
		df	1024	1024	1024	0	1024
	Strength	Correlation	.225	-.272	.116	-.005	1.000
		Significance (2-tailed)	.000	.000	.000	.866	.
		df	1024	1024	1024	1024	0

Based on the partial correlation matrix, the “Water” variable exhibits the highest correlation to the “Strength” with a partial correlation magnitude of 0.272 as compared to other independent variables. In addition, having a significance less than 0.05. Therefore, the “Water” variable will be the next variable to enter the model.

Step 5: Model 5

Variables added from previous steps: “Cement”, “Superplasticizer”, “Age”, “Blast Furnace Slag”, “Water”

To identify the next significant independent variable, the partial correlation is used. Table 2.7 shows the partial correlation matrix after excluding the variable added from previous step.

Table 2.7: Partial correlation matrix for model 5

Correlations

Control Variables			Fly Ash	Coarse Aggregate	Fine Aggregate	Strength
Cement & Superplasticizer & Age & Blast Furnance Slag & Water	Fly Ash	Correlation	1.000	-.125	-.521	.267
		Significance (2-tailed)	.	.000	.000	.000
		df	0	1023	1023	1023
	Coarse Aggregate	Correlation	-.125	1.000	-.637	-.014
		Significance (2-tailed)	.000	.	.000	.653
		df	1023	0	1023	1023
	Fine Aggregate	Correlation	-.521	-.637	1.000	-.126
		Significance (2-tailed)	.000	.000	.	.000
		df	1023	1023	0	1023
	Strength	Correlation	.267	-.014	-.126	1.000
		Significance (2-tailed)	.000	.653	.000	.
		df	1023	1023	1023	0

Based on the partial correlation matrix, the “Fly Ash” variable exhibits the highest correlation to the “Strength” with a partial correlation magnitude of 0.267 as compared to other independent variables. In addition, having a significance less than 0.05. Therefore, the “Fly Ash” variable will be the next variable to enter the model.

Step 6: Model 6

Variables added from previous steps: “Cement”, “Superplasticizer”, “Age”, “Blast Furnace Slag”, “Water”, “Fly Ash”

To identify the next significant independent variable, the partial correlation is used. Table 2.8 shows the partial correlation matrix after excluding the variable added from previous step.

Table 2.8: Partial correlation matrix for model 6

Correlations

Control Variables			Coarse Aggregate	Fine Aggregate	Strength
Cement & Superplasticizer & Age & Blast Furnance Slag & Water & Fly Ash	Coarse Aggregate	Correlation	1.000	-.829	.020
		Significance (2-tailed)	.	.000	.519
		df	0	1022	1022
	Fine Aggregate	Correlation	-.829	1.000	.016
		Significance (2-tailed)	.000	.	.605
		df	1022	0	1022
	Strength	Correlation	.020	.016	1.000
		Significance (2-tailed)	.519	.605	.
		df	1022	1022	0

Based on the partial correlation matrix, the remaining independent variables are exhibiting a significance of greater than 0.05. Therefore, these variables will not be added to the next model. Model 6 will be the final model.

2.2.2 Model Assumption Violation

There are four major assumptions that needs to be satisfied when using multiple linear regression models. This section evaluates whether the model has satisfied the assumptions. The following list the four assumptions:

1) Normality of error terms
2) Linearity of relation between dependent variable and each of independent variables
3) Homoscedasticity of error terms
4) Minimal multicollinearity

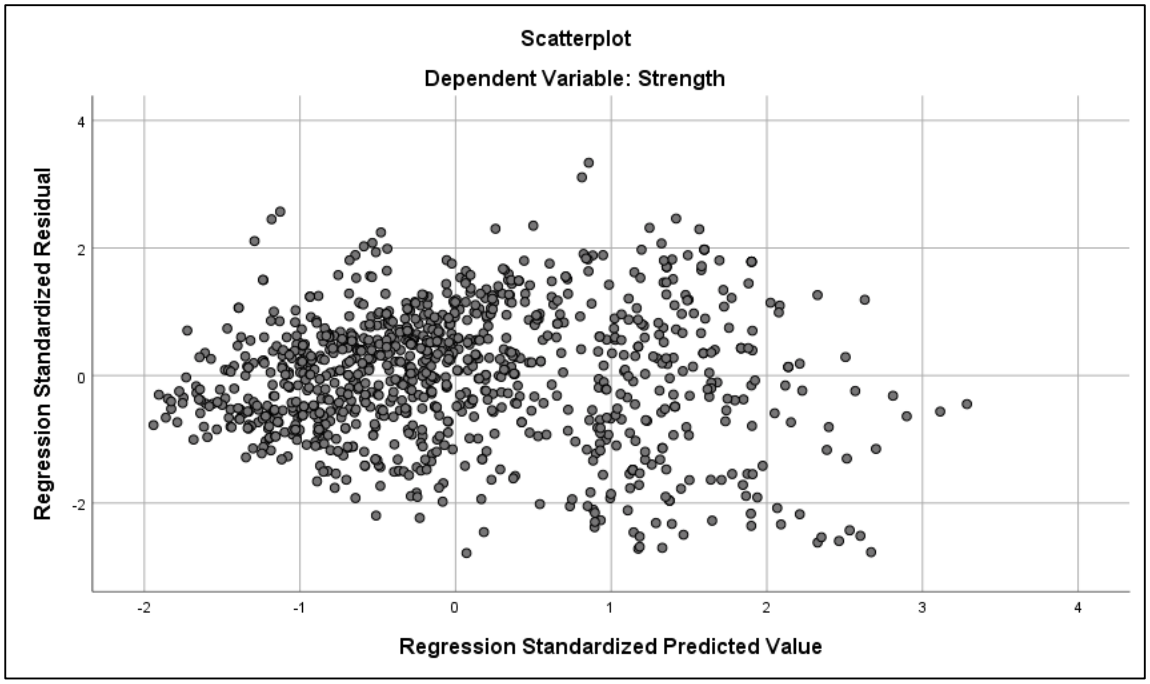

Figure 2.3: Residual plot of model 6

Shown in Figure 2.3, a residual plot for model 6. It is used for evaluating the violation of the first three assumptions of the multiple linear regression. Based on the residual plot, there seems to be no identifiable pattern. This indicates that the error terms are normally distributed and homoscedastic. In addition, it can be concluded that a linear relationship between independent variables and the dependent variable is valid.

Table 2.9: Coefficients table of model 6

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	28.993	4.213		6.881	.000		
	Cement	.105	.004	.659	24.825	.000	.535	1.870
	Blast Furnance Slag	.086	.005	.447	17.385	.000	.572	1.749
	Fly Ash	.069	.008	.263	8.877	.000	.430	2.327
	Water	-.218	.021	-.279	-10.322	.000	.517	1.933
	Superplasticizer	.240	.085	.086	2.842	.005	.413	2.424
	Age	.113	.005	.429	20.988	.000	.902	1.108

a. Dependent Variable: Strength

Table 2.9 shows the coefficient table generated by SPSS of model 6. It is used to evaluate the minimal multicollinearity assumption. Based on the variance inflation factor (VIF), it can be identified that all independent variables have a VIF of less than 10. This indicates there is minimal collinearity.

Based on the evaluation, it can be identified that the model satisfied all four assumptions of multiple linear regression. Thus, there is no violation of assumptions.

2.2.3 Model Summary

Table 2.10 shows the model summary table generated from SPSS. The R^2 value increases as the number of variables are added to the model. Model 6 which is the final model, has a R^2 of 0.614 which is the highest among all other models. This indicates 61.4% of variance in "Strength" can be explained by model 6, which consist of six predictors "Cement", "Superplasticizer", "Age", "Blast Furnace Slag", "Water", and "Fly Ash".

Table 2.10: Model summary table

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.498[a]	.248	.247	14.49549
2	.593[b]	.351	.350	13.46953
3	.694[c]	.482	.480	12.04392
4	.742[d]	.551	.549	11.21432
5	.764[e]	.584	.582	10.79734
6	.784[f]	.614	.612	10.40918

a. Predictors: (Constant), Cement

b. Predictors: (Constant), Cement, Superplasticizer

c. Predictors: (Constant), Cement, Superplasticizer, Age

d. Predictors: (Constant), Cement, Superplasticizer, Age, Blast Furnance Slag

e. Predictors: (Constant), Cement, Superplasticizer, Age, Blast Furnance Slag, Water

f. Predictors: (Constant), Cement, Superplasticizer, Age, Blast Furnance Slag, Water, Fly Ash

2.2.4 Model Adequacy

To identify the model adequacy, the analysis of variance (ANOVA) table can be used. Table 2.11 shows the ANOVA results generated for model 6.

Table 2.11: ANOVA

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	176331.986	6	29388.664	271.235	.000[b]
	Residual	110843.202	1023	108.351		
	Total	287175.187	1029			

a. Dependent Variable: Strength

b. Predictors: (Constant), Age, Blast Furnance Slag, Superplasticizer, Cement, Water, Fly Ash

To test for model adequacy, the following hypothesis test is used:

H_o: The model is not adequate.

H_1: The model is adequate.

Based on Table 2.11, the ANOVA results for model 6 shows a significance of less than 0.05. Therefore, null hypothesis is rejected. This indicates that the model is adequate.

2.2.5 Hypothesis Testing for Model Coefficients

The hypothesis testing for each model coefficient will be based on Table 2.9 which is the coefficients table of model 6.

Hypothesis test for “Cement”:

H_o: “Cement” does not affect “Strength”.

H_1: “Cement” does affect “Strength”.

The significance is less than 0.05 thus null hypothesis is rejected. Therefore, “Cement” does have an effect on “Strength”.

Hypothesis test for “Blast Furnace Slag”:

H_o: “Cement” does not affect “Strength”.

H_1: “Cement” does affect “Strength”.

The significance is less than 0.05 thus null hypothesis is rejected. Therefore, “Blast Furnace Slag” does have an effect on “Strength”.

Hypothesis test for “Fly Ash”:

H_o: “Cement” does not affect “Strength”.

H_1: “Cement” does affect “Strength”.

The significance is less than 0.05 thus null hypothesis is rejected. Therefore, “Fly Ash” does have an effect on “Strength”.

Hypothesis test for “Water”:

H_o: “Cement” does not affect “Strength”.

H_1: “Cement” does affect “Strength”.

The significance is less than 0.05 thus null hypothesis is rejected. Therefore, “Water” does have an effect on “Strength”.

Hypothesis test for “Superplasticizer”:

H_o: “Cement” does not affect “Strength”.

H_1: “Cement” does affect “Strength”.

The significance is less than 0.05 thus null hypothesis is rejected. Therefore, “Superplasticizer” does have an effect on “Strength”.

Hypothesis test for “Age”:

H_o: “Cement” does not affect “Strength”.

H_1: “Cement” does affect “Strength”.

The significance is less than 0.05 thus null hypothesis is rejected. Therefore, “Age” does have an effect on “Strength”.

Therefore, the final model can be represented by the following equation:

$$Strength = 28.993 + 0.105(Cement) + 0.086(Blast\ Furnace\ Slag) + 0.069(Fly\ Ash) - 0.218(Water) + 0.240(Superplasticizer) + 0.113(Age)$$

SECTION 3

CONCLUSION

A multiple linear regression model was developed to predict the concrete compressive strength. The model was adequate and achieved a R^2 of 0.612 but it was lower than the results from literature. The low R^2 can be due to the fitting ability of linear regression models, which are not optimized to fit the non-linear nature of the concrete data. This indicates additional improvement can be done to the model to improve the R^2 value. Suggestion to improve the R^2 would include the use of non-linear models which can better fit the multitude of variables that can affect the concrete compressive strength. In addition, using a bigger dataset would allow the model to better learn the patterns within the data. The objectives of this study were achieved, and it was identified that water, blast furnace slag, and fly ash have an effect on the concrete compressive strength. While the coarse aggregate does not have an effect on the concrete compressive strength.

PART B

FACTOR ANALYSIS

SECTION 4

FACTOR ANALYSIS

4.1 INTRODUCTION

4.1.1 Purpose of Factor Analysis

Factor analysis is an interdependence technique. It refers to the grouping of several variables into few factors. This is done by combining similar variables to form one representative variable. The use of factor analysis reduces the number of variables to be examined which facilitates the simplifying of data to achieve parsimony. In addition, factor analysis facilitates the analysis of large dataset with large number of variables in a condensed manner.

Concrete is a heterogeneous material, meaning the composition of concrete is not uniform throughout which contains a mixture of different materials. This led to the complexity in predicting the concrete compressive strength as there are a few materials involved in the composition of concrete which give rise to the high number of variables. This increases model complexity and may affect the model fitness. The use of factor analysis would facilitate the reduction of number of variables so that a simpler model can be achieved which might lead to improvement of the model prediction capability.

4.1.2 Exclusion of Non-Metric Variables

However, there is a limitation in using factor analysis, which is the type of variables that can be used in factor analysis. Factor analysis underlying workings is based on the correlation among variables. Correlation among metric variables can be easily calculated and several types of correlation measures exists. However, the determination of correlation among non-metric variables are vastly different from the metric variables. The non-metric variables cannot utilize the same type of correlation measures as the metric variables. Therefore, it is not advisable to include non-metric variables into the factor analysis. Although, non-metric variables can be transformed into metric variables using dummy variables. It is not advisable, as the factor analysis cannot ensure that the group of dummy variables derived from the single variable falls under a single factor.

4.2 RESULT AND ANALYSIS

4.2.1 Bartlett's Test of Sphericity

Table 4.1 shows the Bartlett's test of sphericity results generated from the factor analysis. The test is used to determine if the correlation matrix has an identity matrix. The hypothesis for the test is as followed:

H_0: The data is not factorable.

H_1: The data is factorable.

Since the significance obtained is less than 0.05, the null hypothesis is rejected. Thus, the data is factorable.

Table 4.1: Bartlett's Test of Sphericity

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.554
Bartlett's Test of Sphericity	Approx. Chi-Square	1363.869
	df	15
	Sig.	.000

4.2.2 Anti-Image Correlation

Table 4.2 shows the anti-image correlation matrix which contains the measures of sampling adequacy (MSA) for a variable. It can be identified that the selected variables have a measure of sampling adequacy greater than 0.5.

Table 4.2: Anti-image matrices

Anti-image Matrices							
		Age	Blast Furnance Slag	Fly Ash	Water	Superplasticizer	Fine Aggregate
Anti-image Covariance	Age	.898	.111	.108	-.115	-.027	.060
	Blast Furnance Slag	.111	.756	.293	-.050	-.150	.208
	Fly Ash	.108	.293	.716	-.019	-.206	.090
	Water	-.115	-.050	-.019	.453	.282	.206
	Superplasticizer	-.027	-.150	-.206	.282	.484	.014
	Fine Aggregate	.060	.208	.090	.206	.014	.724
Anti-image Correlation	Age	.684[a]	.134	.135	-.180	-.041	.075
	Blast Furnance Slag	.134	.580[a]	.398	-.086	-.249	.281
	Fly Ash	.135	.398	.521[a]	-.034	-.351	.125
	Water	-.180	-.086	-.034	.597[a]	.602	.359
	Superplasticizer	-.041	-.249	-.351	.602	.547[a]	.024
	Fine Aggregate	.075	.281	.125	.359	.024	.612[a]

a. Measures of Sampling Adequacy(MSA)

4.2.3 Communality

Table 4.3 shows the communality table generated from SPSS, which assess how well each variable is explained by the factors. A high communality value would indicate that a variable is well represented by the factors. While a low communality value would indicate that a variable has lesser in common with other variables.

Table 4.3: Communalities

Communalities		
	Initial	Extraction
Age	1.000	.320
Blast Furnance Slag	1.000	.855
Fly Ash	1.000	.426
Water	1.000	.758
Superplasticizer	1.000	.704
Fine Aggregate	1.000	.407
Extraction Method: Principal Component Analysis.		

The following outlines the communality interpretation of each variable:

32.0% of the variance in “Age” is explained by the factors.
85.5% of the variance in “Blast Furnace Slag” is explained by the factors.
42.6% of the variance in “Fly Ash” is explained by the factors.
75.8% of the variance in “Water” is explained by the factors.
70.4% of the variance in “Superplasticizer” is explained by the factors.
40.7% of the variance in “Fine Aggregate” is explained by the factors.

4.2.4 Eigenvalue

Table 4.4 shows the variance explained by each factor which can be determined based on the eigenvalues. The eigenvalue is a measure to identify how much variance a factor can explain. A factor with an eigenvalue greater than one indicates a higher variance is explained as compared to a single variable. Based on the output, component one and component two has an eigenvalue greater than one. This indicates component one would explain as much variance as 2.267 of the three variables. While component two would explain as much variance as 1.204 of the three variables. Three variables from each component are identified based on Table 4.6.

Table 4.4: Total variance explained

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.267	37.785	37.785	2.267	37.785	37.785	1.921	32.025	32.025
2	1.204	20.068	57.853	1.204	20.068	57.853	1.550	25.829	57.853
3	.969	16.153	74.006						
4	.856	14.273	88.279						
5	.434	7.236	95.515						
6	.269	4.485	100.000						

Extraction Method: Principal Component Analysis.

Figure 4.1 shows the scree plot based on the eigenvalues generated. According to the scree plot, the number of components to be retained can be identified as two as it has an eigenvalue of greater than one. In addition, only two components are retained due to the change in eigenvalue after two components become more gradual and lesser.

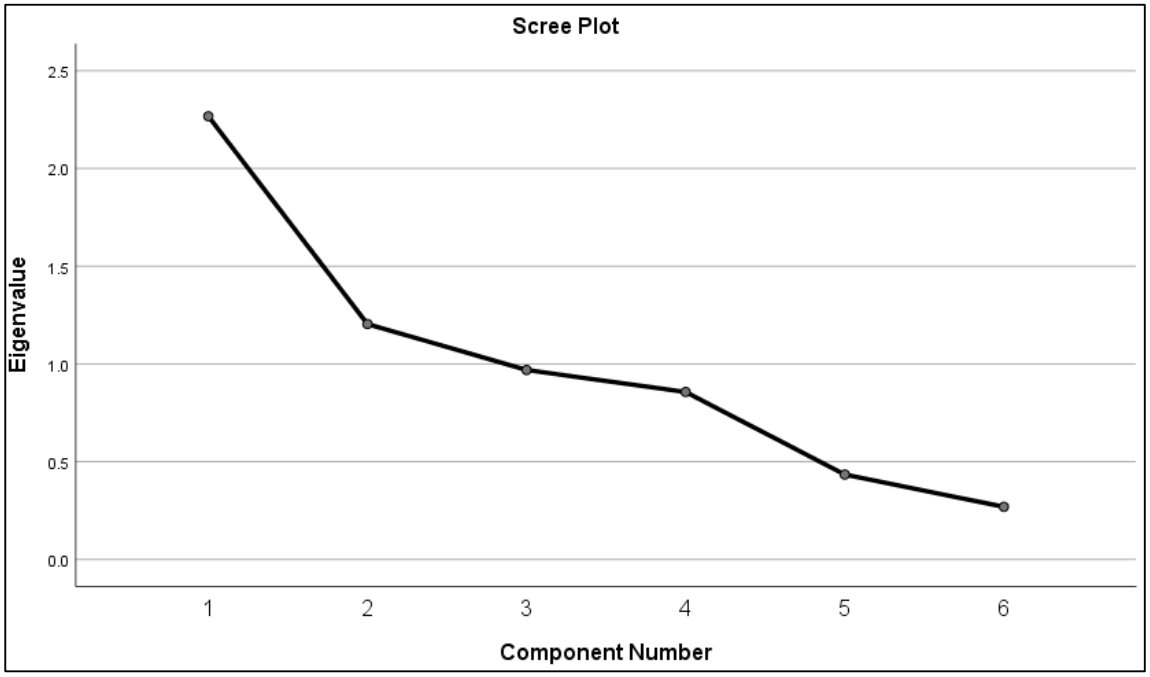

Figure 4.1: Scree plot

4.2.5 Grouping Variables

Table 4.5 shows the rotated component matrix indicating that two components has been identified based on the varimax rotation. The first factor can be identified to consist of variable “Superplasticizer”, “Water”, and “Age”. While the second factor can be identified to consist of variable “Blast Furnace Slag”, “Fly Ash”, and “Fine Aggregate”.

Table 4.5: Rotated component matrix

Rotated Component Matrix[a]

	Component	
	1	2
Superplasticizer	.824	
Water	-.811	
Age	-.564	
Blast Furnance Slag		-.892
Fly Ash		.587
Fine Aggregate		.530

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

4.2.6 Improving Factorability

Factorability refers to variables that have high tendency of being factored. Factorability can be improved by introducing more variables that exhibit higher multicollinearity. In addition, sample size of the dataset can be increased to achieve better correlation estimates among variables. Moreover, the elimination of variables can be done by removing variable with MSA value less than 0.5.

4.2.7 Factor Cross-Loading

Factor cross-loading refers to a variable that has more than one significant loading which indicates that a variable loading is in more than one factor. This would cause complexity in identifying which factor such variable should belong to.

To reduce the effect of factor cross-loading, different rotation methods can be experimented to define a simpler structure. However, if cross-loading persist after rotation, consideration of removal of variable from analysis can be performed. Referring Table 4.5, this analysis adopted the Varimax rotation since it is the commonly used method. This reduced the cross-loading effect resulting in easier identification of variables in different components.

REFERENCES

Chithra, S., Kumar, S. R. R. S., Chinnaraju, K., & Alfin Ashmita, F. (2016). A comparative study on the compressive strength prediction models for High Performance Concrete containing nano silica and copper slag using regression analysis and Artificial Neural Networks. *Construction and Building Materials, 114*, 528-535. doi:https://doi.org/10.1016/j.conbuildmat.2016.03.214

Jin, R., Chen, Q., & Soboyejo, A. B. O. (2018). Non-linear and mixed regression models in predicting sustainable concrete strength. *Construction and Building Materials, 170*, 142-152. doi:https://doi.org/10.1016/j.conbuildmat.2018.03.063

Song, H., Ahmad, A., Farooq, F., Ostrowski, K. A., Maślak, M., Czarnecki, S., et al. (2021). Predicting the compressive strength of concrete with fly ash admixture using machine learning algorithms. *Construction and Building Materials, 308*, 125021. doi:https://doi.org/10.1016/j.conbuildmat.2021.125021

Watts, J. (2019). Concrete: the most destructive material on Earth. *The Guardian*. Retrieved from https://www.theguardian.com/cities/2019/feb/25/concrete-the-most-destructive-material-on-earth

Yeh, I.-C. (1998). Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research, 28*(12), 1797 -1808.

Young, B. A., Hall, A., Pilon, L., Gupta, P., & Sant, G. (2019). Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods. *Cement and Concrete Research, 115*, 379-388. doi:https://doi.org/10.1016/j.cemconres.2018.09.006