Name: Tong Ren
Registration ID: 39899381
EMAIL: tongren1984@gmail.com
Regression Analysis Project
Winter 2012

Introduction

Concrete is the most important material in civil engineering of construction and maintenance of the physical and naturally built environment. The project analyzes the effectiveness of the relationships between ingredients and concrete compressive strength. The Excel Regression Add-In was used to run regression analysis.

Data

Data Source:

Website http://archive.ics.uci.edu/ml/machine-learning-databases/concrete/compressive/Concrete_Data.xls

Description: The data are related to concrete compressive strength depending on age and ingredients. The ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate and fine aggregate. The dependent variable (denoted by Y) is concrete compressive strength.

Number of cases: 1030

Variables

Dependent Variable: Concrete compressive strength -- quantitative -- MPa
Independent Variables:

1. Cement (component 1) -- quantitative – kg/m^3
2. Blast Furnace Slag (component 2) – quantitative – kg/m^3
3. Fly Ash (component 3) -- quantitative – kg/m^3
4. Water (component 4) -- quantitative – kg/m^3
5. Superplasticizer (component 5) – quantitative – kg/m^3
6. Coarse Aggregate (component 6) – quantitative – kg/m^3
7. Fine Aggregate (component 7) – quantitative – kg/m^3
8. Age -- quantitative -- Day (1~365)

Equation and Variables

Where the equation and variables for the full model are:

$$Y= \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7 + \beta_8X_8 + \varepsilon_i$$

Where: Y = Concrete compressive strength -- quantitative -- MPa

α = Intercept
βi = Least squares coefficients
X_1 = Cement (component 1) -- quantitative – kg/m^3
X_2 = Blast Furnace Slag (component 2) – quantitative – kg/m^3
X_3 = Fly Ash (component 3) -- quantitative – kg/m^3
X_4 = Water (component 4) -- quantitative – kg/m^3
X_5 = Superplasticizer (component 5) – quantitative – kg/m^3
X_6 = Coarse Aggregate (component 6) – quantitative – kg/m^3
X_7 = Fine Aggregate (component 7) – quantitative – kg/m^3
X_8 = Age -- quantitative -- Day (1~365)

Hypothesis

The null hypothesis is that all least squares coefficients are zero: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

Data Analysis

Results from the Regression Analysis

The regression was performed for some combinations of the above stated variables. Results are illustrated below separately:

Model 1 (Using All Explanatory Variables)

Regression Statistics	
Multiple R	0.7845156
R Square	0.6154647
Adjusted R Square	0.6124517
Standard Error	10.399849
Observations	1030

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	176744.87	22093.109	204.26914	6.76E-206
Residual	1021	110428.16	108.15686		
Total	1029	287173.03			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-23.163756	26.588421	-0.8711971	0.3838513	-75.337951	29.010439	-75.337951	29.010439
Cement	0.11978	0.0084894	14.1100	1.963E-	0.10312	0.13644	0.103126	0.136443

	53		32	41	67	39	7	9
Blast Furnace Slag	0.1038472	0.0101362	10.245169	1.633E-23	0.0839571	0.1237374	0.0839571	0.1237374
Fly Ash	0.0879431	0.0125851	6.987866	5.03E-12	0.0632474	0.1126387	0.0632474	0.1126387
Water	-0.1502979	0.0401793	-3.7406843	0.0001937	-0.2291413	-0.0714546	-0.2291413	-0.0714546
Superplasticizer	0.2906869	0.0934599	3.110286	0.0019209	0.1072915	0.4740823	0.1072915	0.4740823
Coarse Aggregate	0.0180302	0.0093942	1.9192866	0.0552265	-0.000404	0.0364644	-0.000404	0.0364644
Fine Aggregate	0.0201545	0.0107027	1.883121	0.059968	-0.0008473	0.0411562	-0.0008473	0.0411562
Age	0.1142256	0.0054275	21.045741	5.841E-82	0.1035753	0.1248759	0.1035753	0.1248759

The corresponding regression equation is:

$Y= -23.1638 + 0.1198X_1 + 0.1038X_2 + 0.08794X_3 - 0.1503X_4 + 0.2907X_5 + 0.01803X_6 + 0.02015X_7 + 0.1142X_8$

The adjusted R^2 statistic of 0.61245 means that about 61% of variation in the dependent variable, i.e. concrete compressive strength, can be explained by these eight variables which suggest a very reasonable relationship of these explanatory variables to concrete compressive strength.

The T-statistic for variable X_8 (age) is the highest (21.0457) suggesting a great deal of reliance on this variable.

A high p-value means low relevance of an explanatory variable to the dependent variable. Variable X_7 (fine aggregate) has a comparatively very high p-value and hence seemingly irrelevant.

Model 2 (Using Seven Explanatory Variables; dropping X_7)

Regression Statistics	
Multiple R	0.7836639
R Square	0.6141292
Adjusted R Square	0.6114862
Standard Error	10.412796
Observations	1030

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	176361.33	25194.476	232.36495	2.49E-206
Residual	1022	110811.7	108.42632		

Total	1029	287173.03			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	24.06271	8.8425583	2.7212385	0.006614	6.711056	41.41435	6.711065	41.414355
Cement	0.1060683	0.0043655	24.29676	2.72E-103	0.097501	0.114648	0.097501	0.114634
Blast Furnace Slag	0.087453	0.0051978	16.825022	2.925E-56	0.077253	0.097652	0.077253	0.097652
Fly Ash	0.0693289	0.0077992	8.8891876	2.731E-18	0.054024	0.084633	0.054024	0.084633
Water	-0.2112656	0.023824	-8.867752	3.266E-18	-0.25801	-0.16451	-0.258015	-0.164516
Superplasticizer	0.2627082	0.0923862	2.8435866	0.004598	0.081419	0.443996	0.081419	0.443996
Coarse Aggregate	0.0033576	0.0052545	0.6389878	0.522974	-0.00695	0.013668	-0.006953	0.013668
Age	0.1133479	0.0054142	20.935399	2.873E-81	0.102723	0.123972	0.102723	0.123972

The corresponding regression equation is:

$Y = 24.0627 + 0.1061X_1 + 0.08745X_2 + 0.06933X_3 - 0.2113X_4 + 0.2627X_5 + 0.003358X_6 + 0.1133X_8$

The adjusted R^2 statistic of 0.61149 means that about 61% of variation in the dependent variable, i.e. concrete compressive strength, can be explained by these seven variables which suggest a very reasonable relationship of these explanatory variables to concrete compressive strength.

The T-statistic for variable X_8 (age) is the highest (20.9354) suggesting a great deal of reliance on this variable.

A high p-value means low relevance of an explanatory variable to the dependent variable. Variable X_6 (coarse aggregate) has a comparatively very high p-value and hence seemingly irrelevant.

Model 3 (Using Six Explanatory Variables; dropping X_6)

Regression Statistics	
Multiple R	0.7835656
R Square	0.613975
Adjusted R Square	0.6117109
Standard Error	10.409784
Observations	1030

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	176317.06	29386.177	271.18124	1.78E-207
Residual	1023	110855.97	108.3636		
Total	1029	287173.03			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	29.030224	4.2124756	6.8914877	9.639E-12	20.764144	37.296304	20.764144	37.296304
Cement	0.1054275	0.0042475	24.820784	7.82E-107	0.0970926	0.1137624	0.0970926	0.1137624
Blast Furnace Slag	0.0864936	0.0049748	17.386316	1.61E-59	0.0767316	0.0962556	0.0767316	0.0962556
Fly Ash	0.0687084	0.0077363	8.8812901	2.913E-18	0.0535275	0.0838892	0.0535275	0.0838892
Water	-0.2182923	0.0211282	-10.331809	7.193E-24	-0.2597519	-0.1768328	-0.2597519	-0.1768328
Superplasticizer	0.2390025	0.0845858	2.8255638	0.0048116	0.0730211	0.404984	0.0730211	0.404984
Age	0.1134948	0.0054077	20.987498	1.298E-81	0.1028833	0.1241063	0.1028833	0.1241063

The corresponding regression equation is:
$Y= 29.0302 + 0.1054X_1 + 0.08649X_2 + 0.06871X_3 - 0.2183X_4 + 0.2390X_5 + 0.1135X_8$

The adjusted R^2 of the full model is 61.40% demonstrating that the model a good indicator of concrete compressive strength based on those six variables. The p-values for the remaining explanatory variables are very close to zero, allowing us to reject the null hypothesis.

Conclusion

I began the regression analysis with full model of eight variables: cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate and age. I eliminated coarse aggregate, fine aggregate because they did not appear to be a good fit with the model based on the results of eight-variable regression analysis and six-variable regression analysis.

$Y= 29.0302 + 0.1054X_1 + 0.08649X_2 + 0.06871X_3 - 0.2183X_4 + 0.2390X_5 + 0.1135X_6$

Where: Y = Concrete compressive strength -- quantitative -- MPa
X_1 = Cement (component 1) -- quantitative – kg/m^3
X_2 = Blast Furnace Slag (component 2) – quantitative – kg/m^3

X_3 = Fly Ash (component 3) -- quantitative – kg/m^3
X_4 = Water (component 4) -- quantitative – kg/m^3
X_5 = Superplasticizer (component 5) – quantitative – kg/m^3
X_6 = Age -- quantitative -- Day (1~365)

Based on three regression analysis performed, I determined that the equation with six explanatory variables provides the best fit. F-value also is the highest in this model reflecting the fact that the six-variable model is a best fit. (F-values: six variable model = 271.18124; seven variable model = 232.36495; eight variable model = 204.26914)