Runa Veigas
runaveigas@gmail.com

Stepwise Regression for Cement Data

About the Data

Data set that concerns the hardening of cement. In particular, the researchers were interested in learning how the composition of the cement affected the heat evolved during the hardening of the cement. Therefore, they measured and recorded the following data on 13 batches of cement. Variables of this model were,

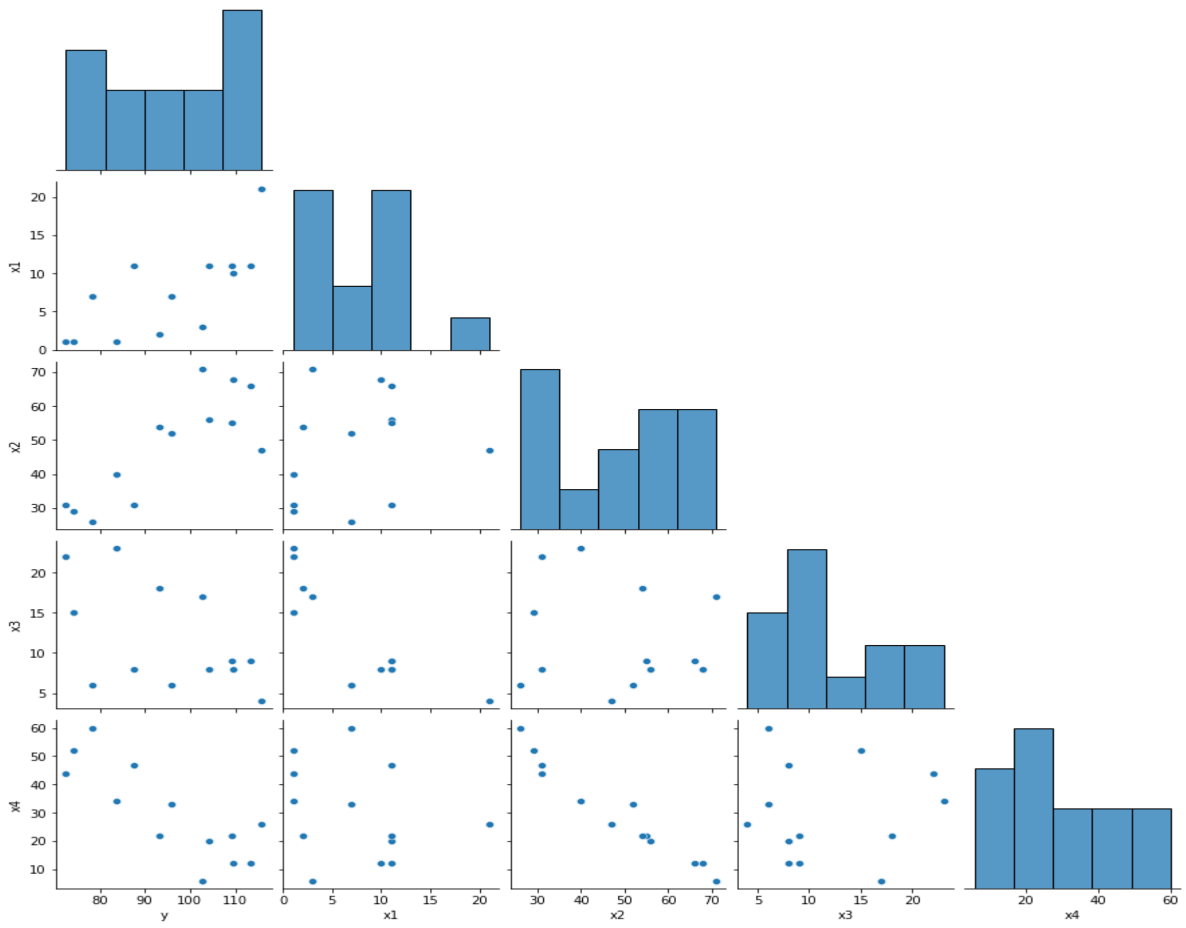
1. Response y: Heat evolved in calories during hardening of cement on a per gram basis
2. Predictor x1: % of tricalcium aluminate
3. Predictor x2: % of tricalcium silicate
4. Predictor x3: % of tetracalcium alumino ferrite
5. Predictor x4: % of dicalcium silicate

Descriptive Statistics

	Y	X1	X2	X3	X4
count	13.000000	13.000000	13.000000	13.000000	13.00000
mean	95.423077	7.461538	48.153846	11.769231	30.00000
count	13	13	13	13	13
min	72.5	1	26	4	6
25%	83.8	2	31	8	20
50%	95.9	7	52	9	26
75%	109.2	11	56	17	44
max	115.9	21	71	23	60

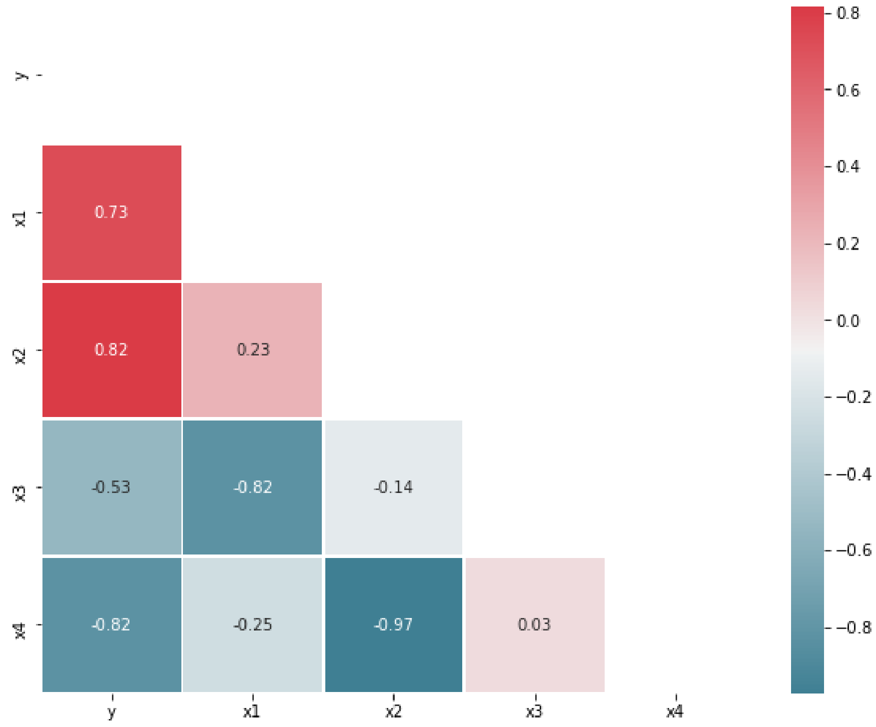
Correlation Analysis

1. **Pairs Plot**



Runa Veigas
runaveigas@gmail.com

2. Correlation Plot



Interpretation: There is a strong positive correlation between y and x_2 and strong negative correlation between y and x_4 and therefore x_2 or x_4 is entered in the model first and considering $\alpha_E = 0.15$ and $\alpha_R = 0.15$.

Stepwise Linear Regression

Regressing *y* on x_1, regressing *y* on x_2, regressing *y* on x_3, and regressing *y* on x_4, we obtain:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	81.4793	4.927	16.536	0.000	70.634	92.324
x1	1.8687	0.526	3.550	0.005	0.710	3.027

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	57.4237	8.491	6.763	0.000	38.736	76.111
x2	0.7891	0.168	4.686	0.001	0.418	1.160

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	110.2027	7.948	13.866	0.000	92.710	127.696
x3	-1.2558	0.598	-2.098	0.060	-2.573	0.061

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	117.5679	5.262	22.342	0.000	105.986	129.150
x4	-0.7382	0.155	-4.775	0.001	-1.078	-0.398

Each of the predictors is a candidate to be entered into the stepwise model because each t-test P-value is less than $\alpha_E = 0.15$. The predictors' x_2 and x_4 tie for having the smallest t-test P-value it is 0.001 in each case. The tie is an artifact of rounding to three decimal places. The t-statistic for x_4 is larger in absolute value than the t-statistic for x_2 is 4.77 versus 4.69 and therefore the P-value for x_4 must be smaller.

As a result of the first step, **we enter x_4 into our stepwise model**. Now we fit each of the two-predictor models that include x_4 as a predictor that is, we regress *y* on x_4 and x_1, regress *y* on x_4 and x_2, and regress *y* on x_4 and x_3, obtaining:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	103.0974	2.124	48.540	0.000	98.365	107.830
x4	-0.6140	0.049	-12.621	0.000	-0.722	-0.506
x1	1.4400	0.138	10.403	0.000	1.132	1.748

Runa Veigas
runaveigas@gmail.com

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	94.1601	56.627	1.663	0.127	-32.013	220.333
x4	-0.4569	0.696	-0.657	0.526	-2.008	1.094
x2	0.3109	0.749	0.415	0.687	-1.357	1.979

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	131.2824	3.275	40.089	0.000	123.986	138.579
x4	-0.7246	0.072	-10.018	0.000	-0.886	-0.563
x3	-1.1999	0.189	-6.348	0.000	-1.621	-0.779

The predictor x_2 is not eligible for entry into the stepwise model because its t-test P-value (0.687) is greater than α_E = 0.15. The predictors x_1 and x_3 are candidates because each t-test P-value is less than α_E = 0.15. The predictors x_1 and x_3 tie for having the smallest t-test P-value < 0.001 in each case. But, again the tie is an artifact of rounding to three decimal places. The t-statistic for x_1 is larger in absolute value than the t-statistic for x_3 10.40 versus 6.35 and therefore the P-value for x_1 must be smaller. **As a result of the second step, we enter x_1 into our stepwise model**.

Now, since x_4 was the first predictor in the model, we must step back and see if entering x_1 into the stepwise model affected the significance of the x_4 predictor. It did not the t-test P-value for testing β_1 = 0 is less than 0.001, and thus smaller than α_R = 0.15. **Therefore, we proceed to the third step with both x_1 and x_4 as predictors in our stepwise model.**

Now, we fit each of the three-predictor models that include x_1 and x_4 as predictors — that is, we regress *y* on x_4, x_1, and x_2, and we regress *y* on x_4, x_1, and x_3, obtaining:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	71.6483	14.142	5.066	0.001	39.656	103.641
x4	-0.2365	0.173	-1.365	0.205	-0.629	0.155
x1	1.4519	0.117	12.410	0.000	1.187	1.717
x2	0.4161	0.186	2.242	0.052	-0.004	0.836

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	111.6844	4.562	24.479	0.000	101.363	122.005
x4	-0.6428	0.045	-14.431	0.000	-0.744	-0.542
x1	1.0519	0.224	4.702	0.001	0.546	1.558
x3	-0.4100	0.199	-2.058	0.070	-0.861	0.041

Both of the remaining predictors x2 and x3 are candidates to be entered into the stepwise model because each t-test P-value is less than α_E = 0.15. The predictor x2 has the smallest t-test P-value (0.052). Therefore, as a result of the third step, we enter x2 into our stepwise model.

Now, since x1 and x4 were the first predictors in the model, we must step back and see if entering x2 into the stepwise model affected the significance of the x1 and x4 predictors. Indeed, it did the t-test P-value for testing $\beta 4$ = 0 is 0.205, which is greater than α_R = 0.15. **Therefore, we remove the predictor x4 from the stepwise model, leaving us with the predictors x1 and x2 in our stepwise model:**

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	52.5773	2.286	22.998	0.000	47.483	57.671
x1	1.4683	0.121	12.105	0.000	1.198	1.739
x2	0.6623	0.046	14.442	0.000	0.560	0.764

Now, we proceed fitting each of the three-predictor models that include x1 and x2 as predictors that is, we regress y on x1, x2, and x3 and we regress y on x1, x2, and x4, obtaining:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	48.1936	3.913	12.315	0.000	39.341	57.046
x1	1.6959	0.205	8.290	0.000	1.233	2.159
x2	0.6569	0.044	14.851	0.000	0.557	0.757
x3	0.2500	0.185	1.354	0.209	-0.168	0.668

Runa Veigas
runaveigas@gmail.com

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	71.6483	14.142	5.066	0.001	39.656	103.641
x1	1.4519	0.117	12.410	0.000	1.187	1.717
x2	0.4161	0.186	2.242	0.052	-0.004	0.836
x4	-0.2365	0.173	-1.365	$\boxed{0.205}$	-0.629	0.155

Neither of the remaining predictors x3 and x4 are eligible for entry into our stepwise model, because each t-test P-value—0.209 and 0.205, respectively is greater than αE = 0.15. That is, we stop our stepwise regression procedure. Our final regression model, based on the stepwise procedure contains only the predictors x1 and x2:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.979
Model:	OLS	Adj. R-squared:	0.974
Method:	Least Squares	F-statistic:	229.5
Date:	Tue, 15 Jun 2021	Prob (F-statistic):	4.41e-09
Time:	15:40:57	Log-Likelihood:	-28.156
No. Observations:	13	AIC:	62.31
Df Residuals:	10	BIC:	64.01
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	52.5773	2.286	22.998	0.000	47.483	57.671
x1	1.4683	0.121	12.105	0.000	1.198	1.739
x2	0.6623	0.046	14.442	0.000	0.560	0.764

Omnibus:	1.509	Durbin-Watson:	1.922
Prob(Omnibus):	0.470	Jarque-Bera (JB):	1.104
Skew:	0.503	Prob(JB):	0.576
Kurtosis:	1.987	Cond. No.	175.

Conclusion

In order to investigate how the composition of the cement affected the heat evolved during the hardening of the cement a Stepwise Regression was carried. The scatter plot showed that there was a strong positive linear relationship between the heat evolved and tri-calcium silicate and heat evolved and di-calcium silicate. Further stepwise regression was conducted to investigate what variables could better predict the heat evolved during harding of the cement. Finally the model obtained was $\hat{y}_i = \beta_o + x_{1i}\beta_1 + x_{2i}\beta_2$ where i = 1,2,3,4. β_o is 52.58, coefficient of x1 is 1.47 and coefficient of x2 is 0.6623. This means that all other variables held constant, for each 1.47 % increase in the tri-calcium aluminate, the model predicts that the heat evolved by the cement increases by 52.58 calories on average and all the variables held constant 0.66 % od increase in tri-calcium silicate, the model predicts that the heat evolved by the cement increases by 52.58 calories on average. The adjusted R2 value obtained was 0.974 which means that 97.4% of the variability in the heat produced can be explained by the model including variables tri-calcium silicate and di-calcium silicate.

************Thank You************