Multiple Linear and Polynomial Regression with Statistical Analysis

Given a set of data of measured (or observed) values of a dependent variable: y_i versus n independent variables x_{1i}, x_{2i}, … x_{ni}, multiple linear regression attempts to find the "best" values of the parameters a_0, a_1, …a_n for the equation

$$\hat{y}_i = a_0 + a_1 x_{1,i} + a_2 x_{2,i} + \ldots + a_n x_{n,i}$$

$\hat{y}_i$ is the calculated value of the dependent variable at point i. The "best" parameters have values that minimize the squares of the errors

$$S = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

In polynomial regression there is only one independent variable, thus

$$\hat{y}_i = a_0 + a_1 x_i + a_2 x_i^2 + \ldots + a_n x_i^n$$

Multiple Linear and Polynomial Regression with Statistical Analysis

Typical examples of multiple linear and polynomial regressions include correlation of temperature dependent physical properties, correlation of heat transfer data using dimensionless groups, correlation of non-ideal phase equilibrium data and correlation of reaction rate data.

The software packages enable high precision correlation of the data, however statistical analysis is essential to determine the ***quality of the fit*** (how well the regression model fits the data) and the ***stability of the model*** (the level of dependence of the model parameters on the particular set of data).

The most important indicators for such studies are the ***residual plot*** (quality of the fit) and ***95% confidence intervals*** (stability of the model)

Regression and Analysis of "Heat of Hardening" Data

No.	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.7
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

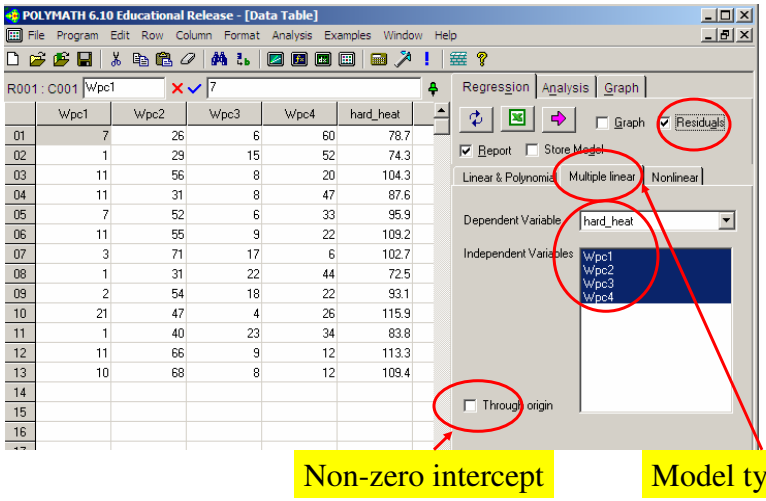
Woods *et al*(1932) investigated the integral heat of hardening of cement as a function of composition. The independent variables represent *weight percent* of the clinker compounds: x_1-tricalcium aluminate ($3CaO \cdot Al2O3$), x_2-tricalcium silicate ($3CaO \cdot SiO2$), x_3-tetracalcium alumino-ferrite ($4CaO \cdot Al2O3 \cdot Fe2O3$), and x_4- β-dicalcium silicate ($3CaO \cdot SiO2$). The dependent variable, y is the total heat evolved (in calories per gram cement) in a 180-day period.

Regression and Analysis of "Heat of Hardening" Data

No.	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.7
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

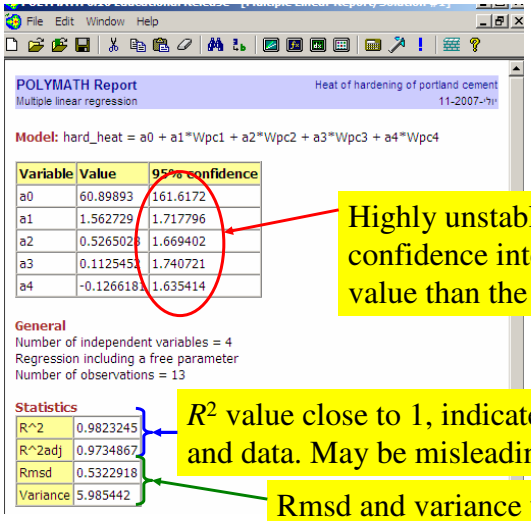
Calculate the coefficients of a linear model representation of y as function of x_1, x_2, x_3, and x_4, calculate the variance and the correlation coefficient R^2 and the 95% confidence intervals.
Prepare a residual plot.
Consider the cases when the model includes and does not include a free parameter.

"Heat of Hardening" Data – Regression and Analysis by Polymath



	Wpc1	Wpc2	Wpc3	Wpc4	hard_heat
01	7	26	6	60	78.7
02	1	29	15	52	74.3
03	11	56	8	20	104.3
04	11	31	8	47	87.6
05	7	52	6	33	95.9
06	11	55	9	22	109.2
07	3	71	17	6	102.7
08	1	31	22	44	72.5
09	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4
14					
15					
16					

"Heat of Hardening" Data –Analysis of the Linear Model that Includes a Free Parameter

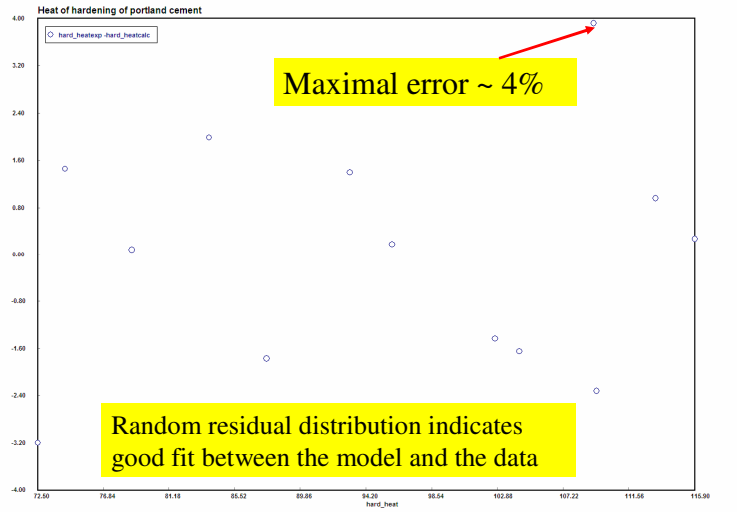


Highly unstable model. All 95% confidence intervals larger in absolute value than the respective parameters.

R^2 value close to 1, indicates good fit between model and data. May be misleading occasionally.

Rmsd and variance values used for comparison between different models

"Heat of Hardening" Data – Residual Plot of the Linear Model that Includes a Free Parameter



"Heat of Hardening" Data – Demonstration of the Harmful Effect of the Instability

Model: hard_heat = a0 + a1*Wpc1 + a2*Wpc2 + a3*Wpc3 + a4*Wpc4

Variable	Value	95% confidence
a0	60.89893	161.6172
a1	1.562729	1.717796
a2	0.526503	1.669402
a3	0.112545	1.740721
a4	-0.12662	1.635414
Statistics		
R^2	0.982325	
Variance	5.985442	

Last data point removed

Variable	Value	95% confidence
a0	36.20362	170.1033
a1	1.783367	1.784379
a2	0.802752	1.770434
a3	0.328399	1.804248
a4	0.12209	1.720564
Statistics		
R^2	0.983897	
Variance	5.746289	

Removal of the last data point causes substantial change in all the parameter values.

In case of the parameter a_4 even its sign is changed

"Heat of Hardening" Data – A Stable Model is Obtained After Removing the Free Parameter

Model: hard_heat = a1*Wpc1 + a2*Wpc2 + a3*Wpc3 + a4*Wpc4

Variable	Value	95% confidence
a1	2.189177	0.4182687
a2	1.154136	0.1082325
a3	0.753295	0.3601112
a4	0.488545	0.093483
Statistics		
R^2	0.980656	
Variance	5.822523	

Last data point removed

Variable	Value	95% confidence
a1	2.151451	0.4078854
a2	1.17869	0.1115862
a3	0.703614	0.3563326
a4	0.487699	0.0901595
Statistics		
R^2	0.983314	
Variance	5.209989	