

Solutions to Practical 9

Question 1

(a) Regression Analysis: Y versus X2

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	434.528	434.528	290.37	0.000
X2	1	434.528	434.528	290.37	0.000
Error	8	11.972	1.496		
Lack-of-Fit	4	3.972	0.993	0.50	0.743
Pure Error	4	8.000	2.000		
Total	9	446.500			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.22329	97.32%	96.98%	95.56%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7.523	0.854	8.81	0.000	
X2	3.932	0.231	17.04	0.000	1.00

Regression Equation

$$Y = 7.523 + 3.932 X_2$$

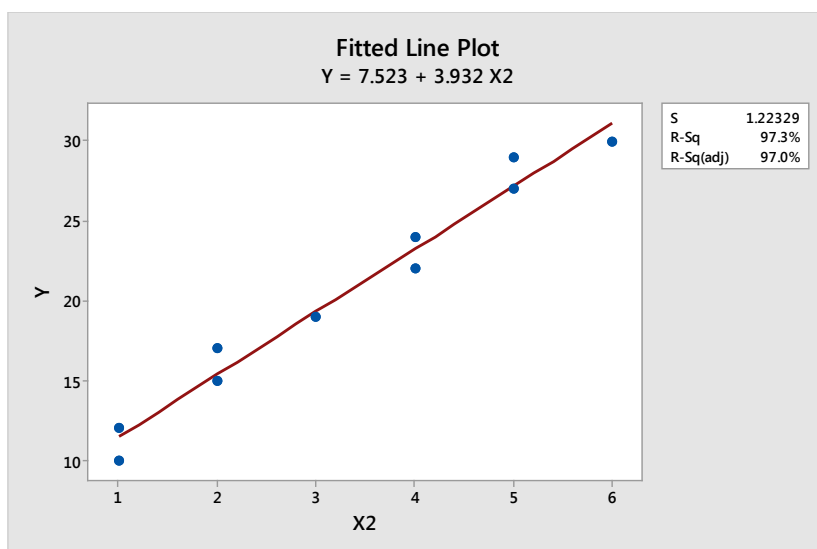
$$H_0 : \beta_2 = 0 \quad H_1 : \beta_2 \neq 0$$

$$\text{Observed Test Statistic} = 17.04 \quad p < 0.001$$

So can reject H_0 in favour H_1 at 1% level : slope is significantly different to zero i.e. a significant relationship

$R^2 = 97.3\%$ so a 'good' linear relationship

$$\text{Fitted Line : } Y = 7.52 + 3.93 X_2$$



(b) Regression Analysis: Y versus X3

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	444.512	444.512	1788.89	0.000
X3	1	444.512	444.512	1788.89	0.000
Error	8	1.988	0.248		
Total	9	446.500			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.498483	99.55%	99.50%	99.16%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	33.267	0.341	97.69	0.000	
X3	-2.3212	0.0549	-42.30	0.000	1.00

Regression Equation

$$Y = 33.267 - 2.3212 X3$$

$$H_0 : \beta_3 = 0 \quad H_1 : \beta_3 \neq 0$$

$$\text{Observed Test Statistic} = -42.30 \quad p < 0.001$$

So can reject H_0 in favour H_1 at 1% level : slope is significantly different to zero i.e. a significant relationship

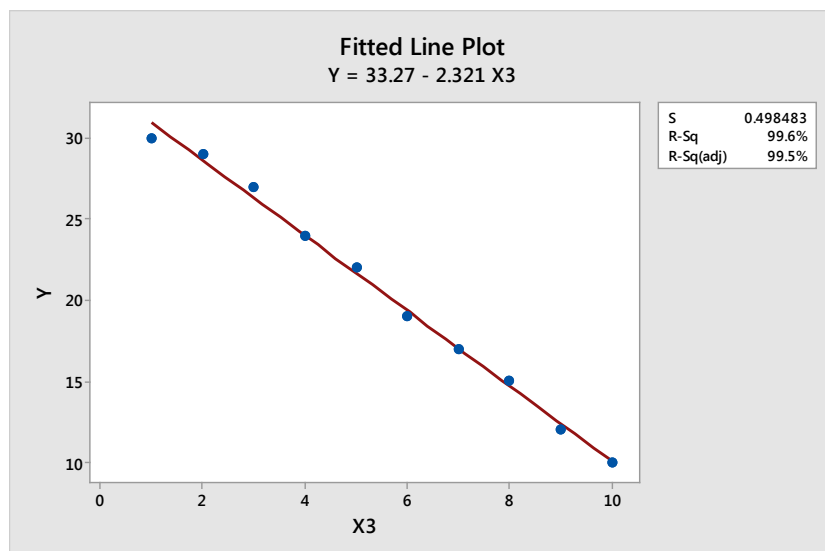
$R^2 = 99.6\%$ so a 'good' linear relationship

$$\text{Fitted Line : } Y = 33.3 - 2.32 X3$$

Fits and Diagnostics for Unusual Observations

Obs	Y	Fit	Resid	Std Resid
10	30.000	30.945	-0.945	-2.34

R Large residual



Y vs X3 is the 'better' relationship since higher R^2

So the better 1 variable model is $Y = 33.3 - 2.32 X3$

Note that I should also have checked the assumptions – you must always do this!! Results are only valid if the assumptions are valid.

(C) Regression Analysis: Y versus X2, X3

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	444.590	222.295	814.83	0.000
X2	1	0.078	0.078	0.29	0.609
X3	1	10.062	10.062	36.88	0.001
Error	7	1.910	0.273		
Total	9	446.500			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.522313	99.57%	99.45%	99.04%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	31.18	3.91	7.97	0.000	
X2	0.323	0.603	0.54	0.609	37.39
X3	-2.135	0.352	-6.07	0.001	37.39

$$H_0 : y = \bar{y} \quad H_1 : y = \alpha + \beta_2 X_2 + \beta_3 X_3$$

Observed Test Statistic = 814.83 p<0.001

So can reject Ho in favour H1 at 1% level so significant regression model

$R^2 = 99.6\%$ so a 'good' linear relationship

$$H_0 : \beta_2 = 0 \quad H_1 : \beta_2 \neq 0$$

Observed Test Statistic = 0.54 p=0.609

So cannot reject Ho in favour H1 at 5% level : slope with X2 is not significantly different to zero i.e. adding X2 to the model once X3 controlled for does not significantly improve the model

$$H_0 : \beta_3 = 0 \quad H_1 : \beta_3 \neq 0$$

Observed Test Statistic = -6.07 p=0.001

So can reject Ho in favour H1 at 5% level : slope with X3 is significantly different to zero i.e. adding X3 to the model once X2 controlled for does significantly improve the model

Regression Equation

$$Y = 31.18 + 0.323 X_2 - 2.135 X_3$$

$$\text{Fitted Line : } Y = 31.2 + 0.323 X_2 - 2.14 X_3$$

Fits and Diagnostics for Unusual Observations

Obs	Y	Fit	Resid	Std Resid
10	30.000	30.981	-0.981	-2.35

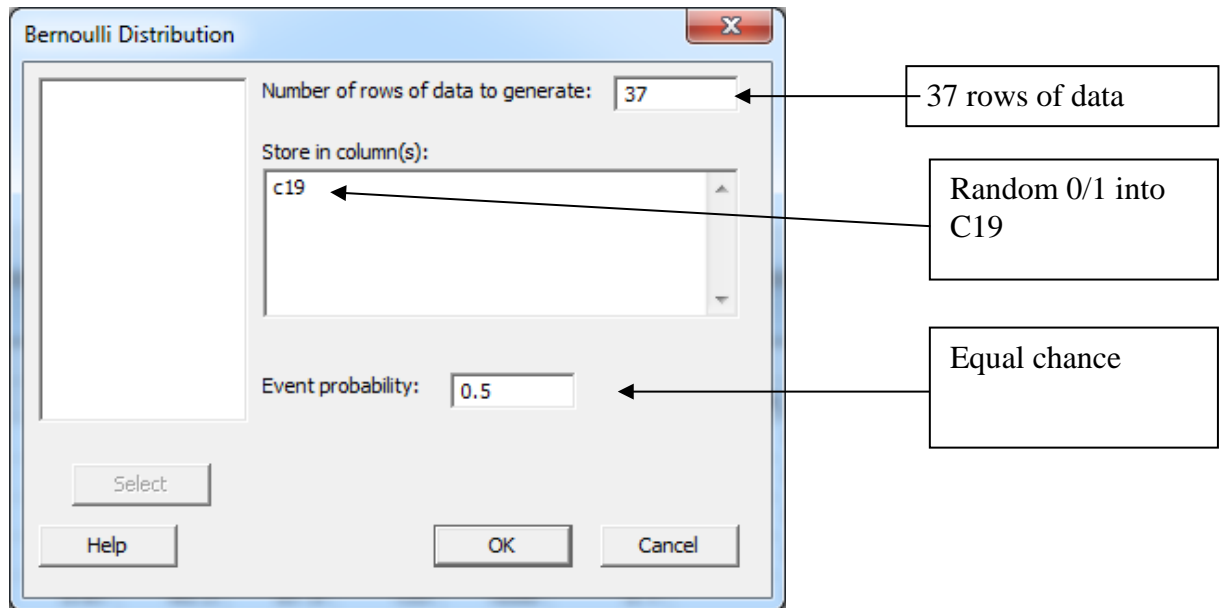
R Large residual

Since adding X2 does not significantly improve the model then the 'best' model for Y is

$$Y = 33.3 - 2.32 X_3$$

Question 2

- (i) Split the data into a **Training Set** and a **Test Set** (with equal chance) using Calc -> Random -> Bernoulli



My random split gave 15 0's and 22 1's so I will use the 1's as the TRAINING SET

(ii) Using Stepwise Regression to find the Best Model for Aroma

Regression Analysis: Aroma_1 versus Cd_1, Mo_1, Mn_1, Ni_1, Cu_1, Al_1, ...

Stepwise Selection of Terms

α to enter = 0.15, α to remove = 0.15

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	20.103	6.7011	19.70	0.000
Cu_1	1	1.100	1.1001	3.23	0.089
Ba_1	1	2.432	2.4321	7.15	0.015
Sr_1	1	12.561	12.5609	36.93	0.000
Error	18	6.122	0.3401		
Total	21	26.226			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.583214	76.65%	72.76%	67.68%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.869	0.348	19.73	0.000	
Cu_1	-1.389	0.772	-1.80	0.089	1.04
Ba_1	6.23	2.33	2.67	0.015	3.39
Sr_1	-3.936	0.648	-6.08	0.000	3.32

Regression Equation

Aroma_1 = 6.869 - 1.389 Cu_1 + 6.23 Ba_1 - 3.936 Sr_1

Fits and Diagnostics for Unusual Observations

Obs	Aroma_1	Fit	Resid	Std Resid
1	3.300	3.285	0.015	0.07 X

X Unusual X

This gives the final Best Model using the method as

Aroma_1 = 6.869 - 1.389 Cu_1 + 6.23 Ba_1 - 3.936 Sr_1

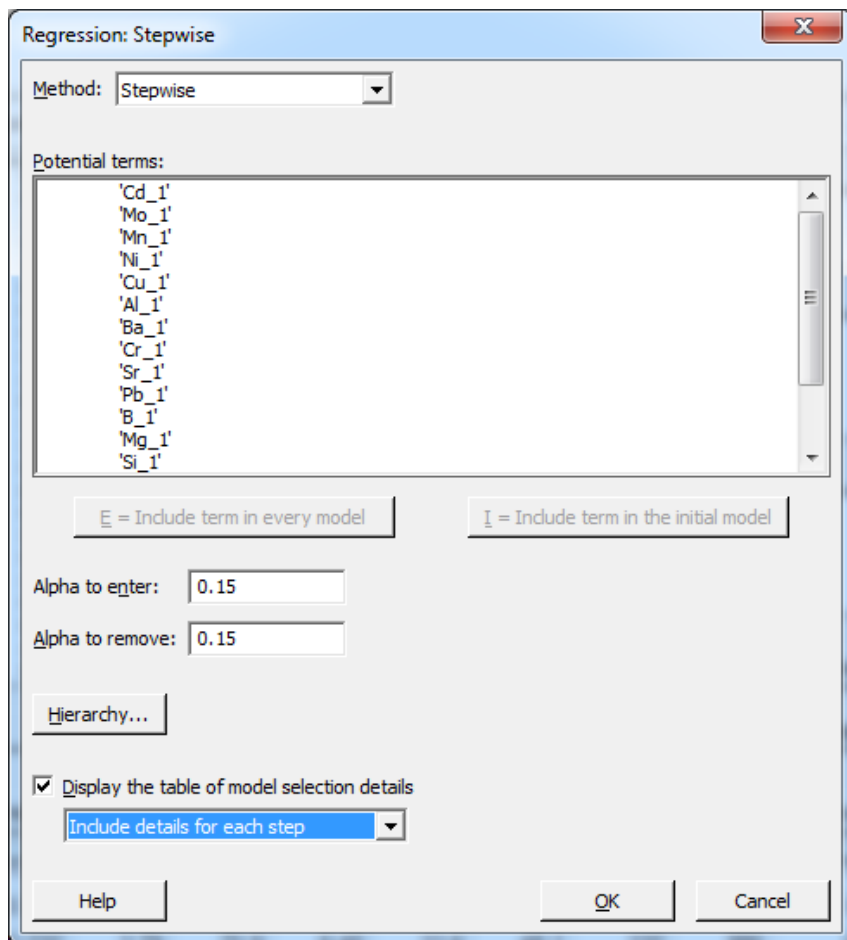
Best Model includes Sr_1, Ba_1 and Cu_1 with $R^2 = 76.65\%$

Note that the slope with Cu_1 is not significantly different to zero at 5% level ($p=0.089>0.05$) i.e. Cu_1 does not significantly improve the model

So could, with justification, use only Sr_1 and Ba_1.

I will, however, use the three variable model including Sr_1, Ba_1 and Cu_1.

If you want to see all the steps in the analysis as we did in lectures the chose 'Include details of each step in the dialogue box below.



The image shows a 'Regression: Stepwise' dialog box. At the top, the 'Method' is set to 'Stepwise'. Below this, a list of 'Potential terms' is shown, including 'Cd_1', 'Mo_1', 'Mn_1', 'Ni_1', 'Cu_1', 'Al_1', 'Ba_1', 'Cr_1', 'Sr_1', 'Pb_1', 'B_1', 'Mg_1', and 'Si_1'. There are two buttons: 'E = Include term in every model' and 'I = Include term in the initial model'. Below these, 'Alpha to enter' and 'Alpha to remove' are both set to 0.15. There is a 'Hierarchy...' button. A checkbox 'Display the table of model selection details' is checked, and a dropdown menu below it is set to 'Include details for each step'. At the bottom are 'Help', 'OK', and 'Cancel' buttons.

Regression: Stepwise

Method: Stepwise

Potential terms:

- 'Cd_1'
- 'Mo_1'
- 'Mn_1'
- 'Ni_1'
- 'Cu_1'
- 'Al_1'
- 'Ba_1'
- 'Cr_1'
- 'Sr_1'
- 'Pb_1'
- 'B_1'
- 'Mg_1'
- 'Si_1'

E = Include term in every model I = Include term in the initial model

Alpha to enter: 0.15

Alpha to remove: 0.15

Hierarchy...

☒ Display the table of model selection details

Include details for each step

Help OK Cancel

The output is on the next page.

Regression Analysis: Aroma_1 versus Cd_1, Mo_1, Mn_1, Ni_1, Cu_1, Al_1, ...

Stepwise Selection of Terms

Candidate terms: Cd_1, Mo_1, Mn_1, Ni_1, Cu_1, Al_1, Ba_1, Cr_1, Sr_1, Pb_1, B_1, Mg_1, Si_1, Na_1, Ca_1, P_1, K_1

	-----Step 1-----		-----Step 2-----		-----Step 3-----	
	Coef	P	Coef	P	Coef	P
Constant	6.963		6.741		6.869	
Sr_1	-2.518	0.000	-3.795	0.000	-3.936	0.000
Ba_1			5.46	0.036	6.23	0.015
Cu_1					-1.389	0.089
S	0.676632		0.616553		0.583214	
R-sq	65.09%		72.46%		76.65%	
R-sq(adj)	63.34%		69.56%		72.76%	
R-sq(pred)	56.02%		62.48%		67.68%	
Mallows' Cp	28.86		20.97		17.34	

α to enter = 0.15, α to remove = 0.15

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	20.103	6.7011	19.70	0.000
Cu_1	1	1.100	1.1001	3.23	0.089
Ba_1	1	2.432	2.4321	7.15	0.015
Sr_1	1	12.561	12.5609	36.93	0.000
Error	18	6.122	0.3401		
Total	21	26.226			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.583214	76.65%	72.76%	67.68%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.869	0.348	19.73	0.000	
Cu_1	-1.389	0.772	-1.80	0.089	1.04
Ba_1	6.23	2.33	2.67	0.015	3.39
Sr_1	-3.936	0.648	-6.08	0.000	3.32

Regression Equation

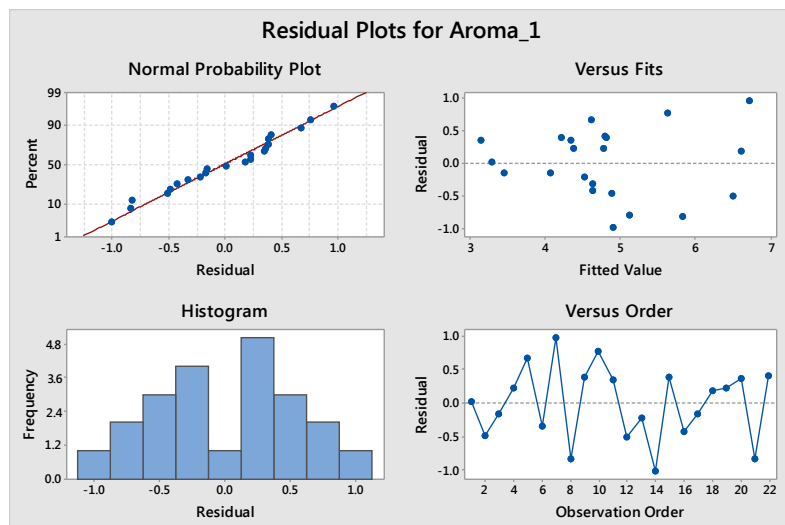
Aroma_1 = 6.869 - 1.389 Cu_1 + 6.23 Ba_1 - 3.936 Sr_1

Fits and Diagnostics for Unusual Observations

Obs	Aroma_1	Fit	Resid	Std Resid	
1	3.300	3.285	0.015	0.07	X

X Unusual X

We must check the assumptions of Normality and Constant Variance for this best model.



Validate Assumptions

Top Left : Can assume normality since graph is approximately linear.

Top Right : Slight problem with assumption of constant variance as points not evenly spread about zero

Regression Analysis: Aroma_1 versus Sr_1, Ba_1, Cu_1

Regression Analysis: Aroma_1 versus Cu_1, Ba_1, Sr_1

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	20.103	6.7011	19.70	0.000
Cu_1	1	1.100	1.1001	3.23	0.089
Ba_1	1	2.432	2.4321	7.15	0.015
Sr_1	1	12.561	12.5609	36.93	0.000
Error	18	6.122	0.3401		
Total	21	26.226			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.583214	76.65%	72.76%	67.68%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.869	0.348	19.73	0.000	
Cu_1	-1.389	0.772	-1.80	0.089	1.04
Ba_1	6.23	2.33	2.67	0.015	3.39
Sr_1	-3.936	0.648	-6.08	0.000	3.32

Regression Equation

$$\text{Aroma}_1 = 6.869 - 1.389 \text{ Cu}_1 + 6.23 \text{ Ba}_1 - 3.936 \text{ Sr}_1$$

Fits and Diagnostics for Unusual Observations

Obs	Aroma_1	Fit	Resid	Std Resid
1	3.300	3.285	0.015	0.07

X Unusual X

Best Model : which explains 76.7% of the variability in Aroma Scores is

$$\text{Aroma}_1 = 6.869 - 1.389 \text{ Cu}_1 + 6.23 \text{ Ba}_1 - 3.936 \text{ Sr}_1$$

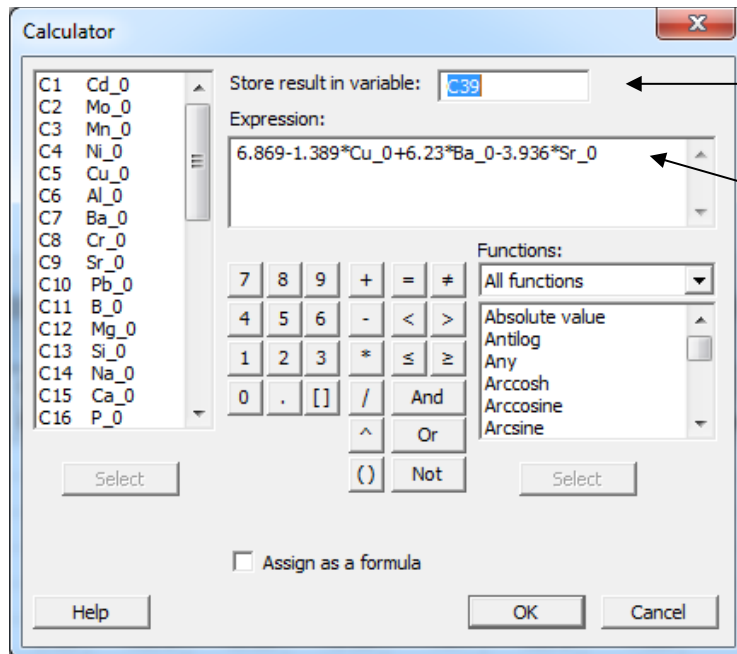
So based on the **Training Set the Best Model for Aroma Scores** using Stepwise Regression is

$$\text{Aroma}_1 = 6.869 - 1.389 \text{ Cu}_1 + 6.23 \text{ Ba}_1 - 3.936 \text{ Sr}_1$$

We must now see how well this works on the Test Set.

- (i) Calculate the Predicted Aroma Scores for the Test Set (i.e. the 0's)

Use Calc -> Calculator

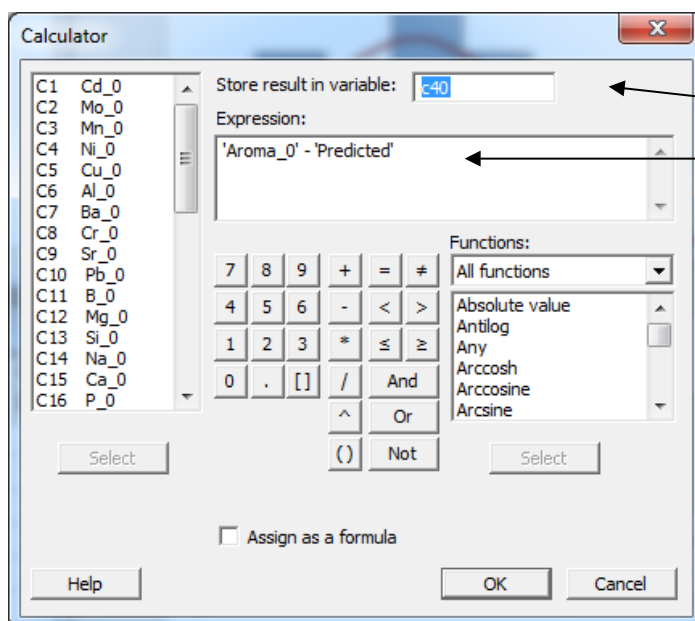


Store Predicted Values in C39

Model is :

Predicted =
$$6.869 - 1.389 * \text{Cu}_0 + 6.23 * \text{Ba}_0 - 3.936 * \text{Sr}_0$$

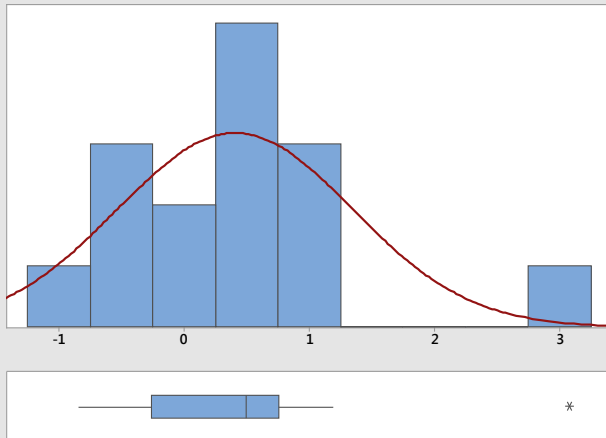
- (ii) Compare the Predicted Aroma Scores (C39) with the Observed Aroma Scores (C18) for the Test Set. To calculate the Prediction Errors - Use Calc -> Calculator ->



Store Prediction Errors (PE) in C40

$$\text{PE} = \text{Observed} - \text{Predicted}$$

Summary Report for PE



Anderson-Darling Normality Test

A-Squared 0.65
P-Value 0.073

Mean 0.40755
StDev 0.93762
Variance 0.87914
Skewness 1.60257
Kurtosis 4.10215
N 15

Minimum -0.84033
1st Quartile -0.26417
Median 0.49233
3rd Quartile 0.75494
Maximum 3.08160

95% Confidence Interval for Mean

-0.11169 0.92679

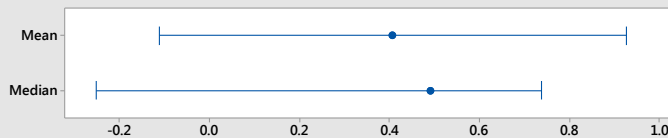
95% Confidence Interval for Median

-0.25068 0.73930

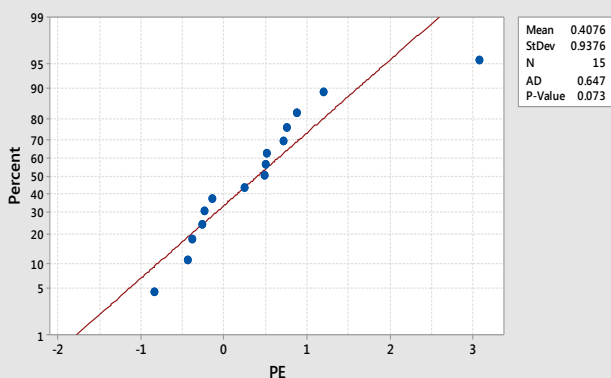
95% Confidence Interval for StDev

0.68646 1.47873

95% Confidence Intervals



Probability Plot of PE



Discussion

Prediction Errors are Normally Distributed (Normal probability plot is approximately linear)

Paired t-test ($p=0.113$) shows we cannot reject H_0 : Mean Difference = 0 in favour of H_1 : Mean difference doesn't equal zero at 5% so on average we have a good model

However, in one instance, we have a prediction error of 3.082. So not always very accurate.

Paired T-Test and CI: Aroma_0, Predicted

Paired T for Aroma_0 - Predicted

	N	Mean	StDev	SE Mean
Aroma_0	15	4.993	1.024	0.264
Predicted	15	4.586	1.623	0.419
Difference	15	0.408	0.938	0.242

95% CI for mean difference: (-0.112, 0.927)

T-Test of mean difference = 0 (vs \neq 0): T-Value = 1.68 P-Value = 0.114

FINDING THE BEST MODEL USING FORWARD STEPPING

Using the same Training and Test Sets.

Regression Analysis: Aroma_1 versus Cd_1, Mo_1, Mn_1, Ni_1, Cu_1, Al_1, ...

Forward Selection of Terms

α to enter = 0.25

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	21.9117	3.6519	12.70	0.000
Cd_1	1	0.6602	0.6602	2.30	0.151
Mo_1	1	1.2084	1.2084	4.20	0.058
Cu_1	1	1.1914	1.1914	4.14	0.060
Ba_1	1	1.9169	1.9169	6.66	0.021
Sr_1	1	7.9071	7.9071	27.49	0.000
Ca_1	1	1.0539	1.0539	3.66	0.075
Error	15	4.3142	0.2876		
Total	21	26.2259			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.536297	83.55%	76.97%	60.39%

Coefficients

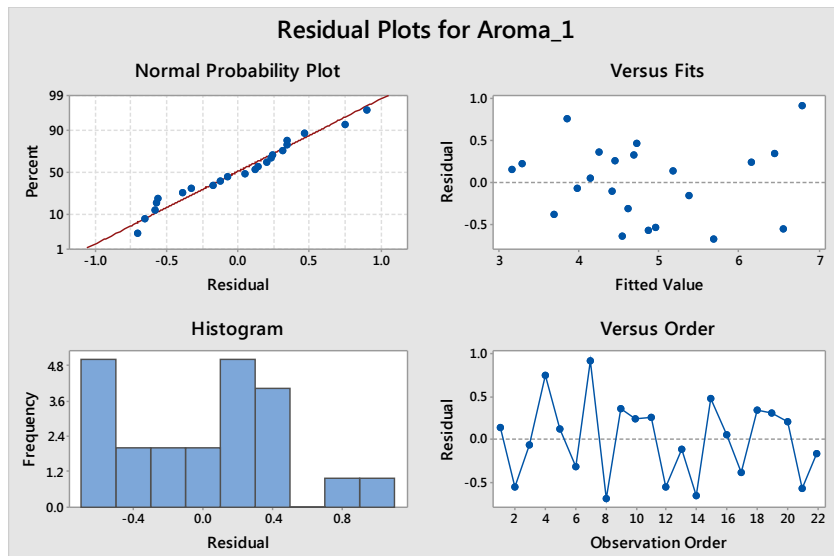
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7.499	0.596	12.59	0.000	
Cd_1	-11.56	7.63	-1.52	0.151	2.92
Mo_1	4.50	2.19	2.05	0.058	2.39
Cu_1	-1.757	0.863	-2.04	0.060	1.53
Ba_1	5.70	2.21	2.58	0.021	3.61
Sr_1	-3.420	0.652	-5.24	0.000	3.98
Ca_1	-0.01274	0.00666	-1.91	0.075	1.32

Regression Equation

Aroma_1 = 7.499 - 11.56 Cd_1 + 4.50 Mo_1 - 1.757 Cu_1 + 5.70 Ba_1 - 3.420 Sr_1 - 0.01274 Ca_1

The Best Model includes Sr_1, Ba_1, Cu_1, Ca_1, Mo_1 and Cd_1

R^2 is increased to 83.55% (and Adjusted R^2 to 76.97%) using this method (again not all slopes are significantly different to zero) but I will proceed using all these variables.



Validate Assumptions

Top Left : Can assume normality since graph is approximately linear.

Top Right : Can assume constant variance as points approx. evenly spread about zero

Regression Analysis: Aroma_1 versus Cd_1, Mo_1, Cu_1, Ba_1, Sr_1, Ca_1

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	21.9117	3.6519	12.70	0.000
Cd_1	1	0.6602	0.6602	2.30	0.151
Mo_1	1	1.2084	1.2084	4.20	0.058
Cu_1	1	1.1914	1.1914	4.14	0.060
Ba_1	1	1.9169	1.9169	6.66	0.021
Sr_1	1	7.9071	7.9071	27.49	0.000
Ca_1	1	1.0539	1.0539	3.66	0.075
Error	15	4.3142	0.2876		
Total	21	26.2259			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.536297	83.55%	76.97%	60.39%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7.499	0.596	12.59	0.000	
Cd_1	-11.56	7.63	-1.52	0.151	2.92
Mo_1	4.50	2.19	2.05	0.058	2.39
Cu_1	-1.757	0.863	-2.04	0.060	1.53
Ba_1	5.70	2.21	2.58	0.021	3.61
Sr_1	-3.420	0.652	-5.24	0.000	3.98
Ca_1	-0.01274	0.00666	-1.91	0.075	1.32

Regression Equation

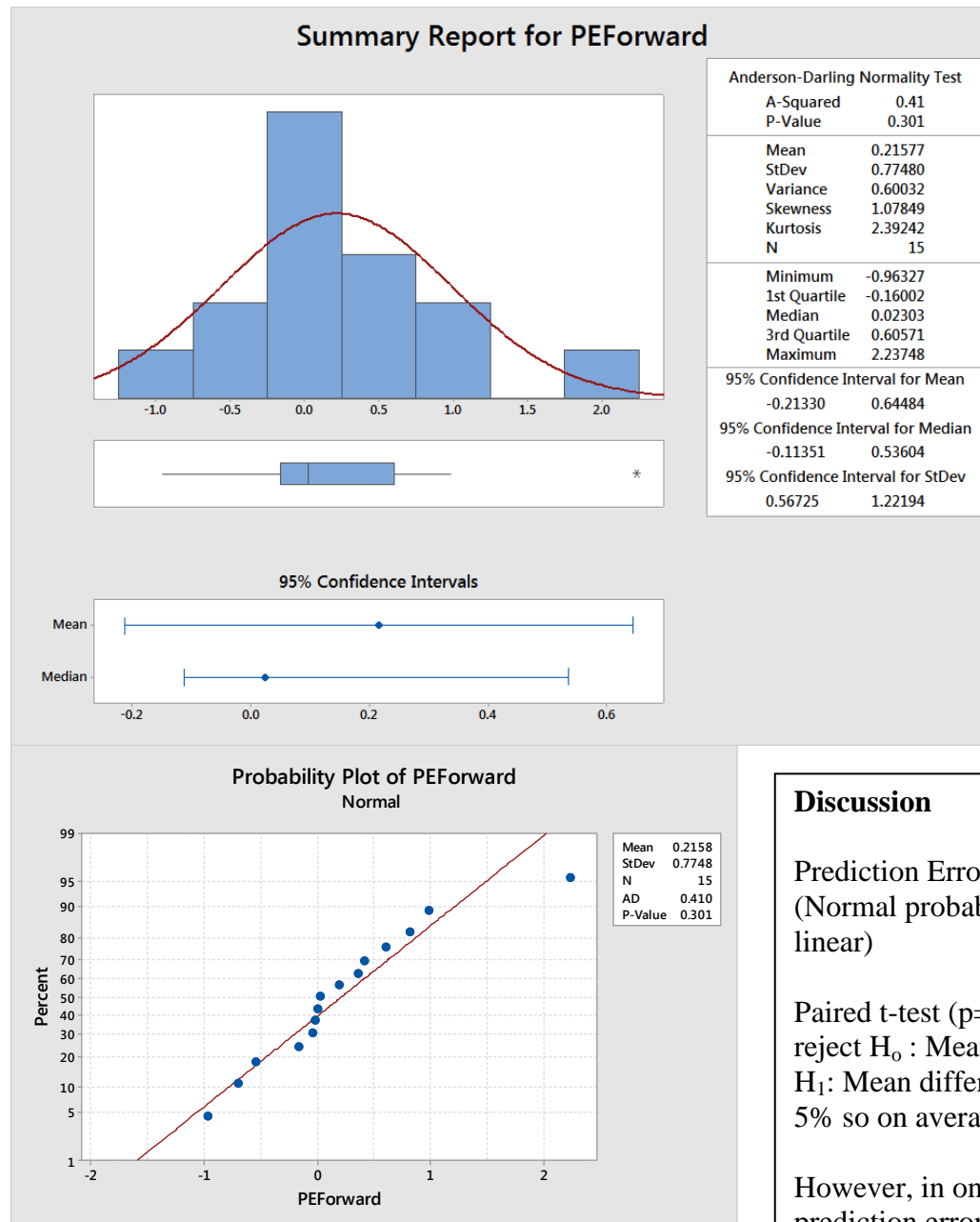
$$\text{Aroma}_1 = 7.499 - 11.56 \text{ Cd}_1 + 4.50 \text{ Mo}_1 - 1.757 \text{ Cu}_1 + 5.70 \text{ Ba}_1 - 3.420 \text{ Sr}_1 - 0.01274 \text{ Ca}_1$$

So Best Model for the Training Set using Forward Stepping is

$$\text{Aroma}_1 = 7.499 - 11.56 \text{ Cd}_1 + 4.50 \text{ Mo}_1 - 1.757 \text{ Cu}_1 + 5.70 \text{ Ba}_1 - 3.420 \text{ Sr}_1 - 0.01274 \text{ Ca}_1$$

Predict the Aroma Scores for the Test Set Model is

Predicted = $7.499 - 11.56 \cdot \text{Cd}_0 + 4.50 \cdot \text{Mo}_0 - 1.757 \cdot \text{Cu}_0 + 5.70 \cdot \text{Ba}_0 - 3.420 \cdot \text{Sr}_0 - 0.01274 \cdot \text{Ca}_0$



Discussion

Prediction Errors are Normally Distributed (Normal probability plot is approximately linear)

Paired t-test ($p=0.299$) shows we cannot reject H_0 : Mean Difference = 0 in favour of H_1 : Mean difference doesn't equal zero at 5% so on average we have a good model

However, in one instance, we have a prediction error of 2.237. So not always very accurate.

Paired T-Test and CI: Aroma_0, PredForward

Paired T for Aroma_0 - PredForward

	N	Mean	StDev	SE Mean
Aroma_0	15	4.993	1.024	0.264
PredForward	15	4.778	1.458	0.377
Difference	15	0.216	0.775	0.200

95% CI for mean difference: (-0.213, 0.645)

T-Test of mean difference = 0 (vs \neq 0): T-Value = 1.08 P-Value = 0.299

FINDING THE BEST MODEL USING BACKWARD STEPPING

Using the same Training and Test Sets.

Regression Analysis: Aroma_1 versus Cd_1, Mo_1, Mn_1, Ni_1, Cu_1, Al_1, ...

Backward Elimination of Terms

α to remove = 0.1

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	11	24.7018	2.2456	14.73	0.000
Cd_1	1	1.4278	1.4278	9.37	0.012
Ni_1	1	1.1467	1.1467	7.52	0.021
Cu_1	1	0.9327	0.9327	6.12	0.033
Ba_1	1	5.3798	5.3798	35.30	0.000
Cr_1	1	5.8275	5.8275	38.24	0.000
Pb_1	1	7.5862	7.5862	49.78	0.000
B_1	1	3.3957	3.3957	22.28	0.001
Na_1	1	2.0367	2.0367	13.36	0.004
Ca_1	1	4.6883	4.6883	30.76	0.000
P_1	1	0.5792	0.5792	3.80	0.080
K_1	1	1.6834	1.6834	11.05	0.008
Error	10	1.5241	0.1524		
Total	21	26.2259			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.390396	94.19%	87.80%	24.58%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.08	1.10	4.60	0.001	
Cd_1	-27.50	8.98	-3.06	0.012	7.63
Ni_1	3.70	1.35	2.74	0.021	2.79
Cu_1	-1.821	0.736	-2.47	0.033	2.10
Ba_1	-9.60	1.62	-5.94	0.000	3.64
Cr_1	-64.5	10.4	-6.18	0.000	3.51
Pb_1	3.748	0.531	7.06	0.000	6.81
B_1	0.3272	0.0693	4.72	0.001	2.37
Na_1	0.01958	0.00536	3.66	0.004	2.54
Ca_1	-0.04449	0.00802	-5.55	0.000	3.63
P_1	-0.00995	0.00510	-1.95	0.080	3.45
K_1	0.00494	0.00149	3.32	0.008	6.67

Regression Equation

$$\begin{aligned} \text{Aroma}_1 = & 5.08 - 27.50 \text{ Cd}_1 + 3.70 \text{ Ni}_1 - 1.821 \text{ Cu}_1 - 9.60 \text{ Ba}_1 - 64.5 \text{ Cr}_1 \\ & + 3.748 \text{ Pb}_1 + 0.3272 \text{ B}_1 + 0.01958 \text{ Na}_1 - 0.04449 \text{ Ca}_1 - 0.00995 \text{ P}_1 \\ & + 0.00494 \text{ K}_1 \end{aligned}$$

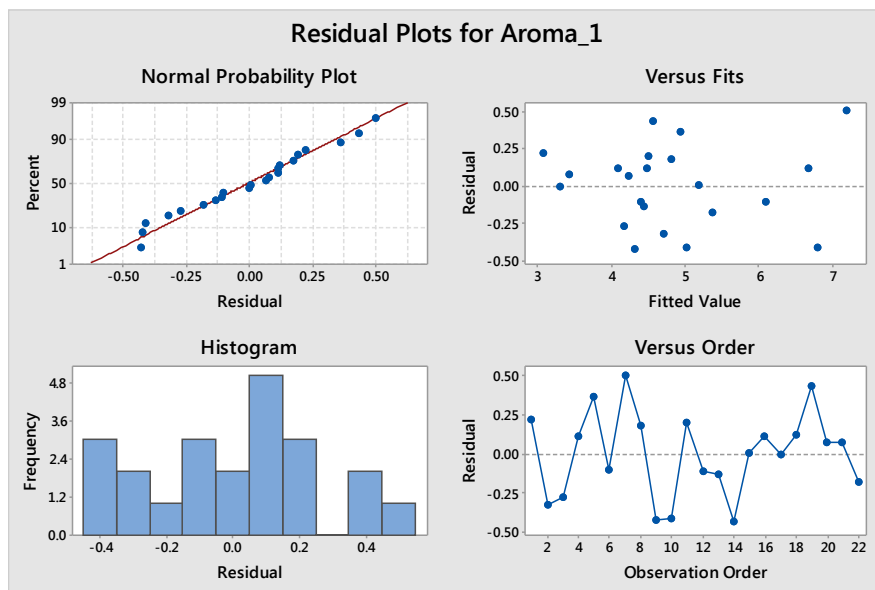
Fits and Diagnostics for Unusual Observations

Obs	Aroma_1	Fit	Resid	Std Resid	
1	3.300	3.078	0.222	2.18	R
7	7.700	7.198	0.502	2.04	R

R Large residual

The Best Model includes Cd_1, Ni_1, Cu_1, Ba_1, Cr_1, Pb_1, B_1, Na_1, Ca_1, P_1 and K_1. (i.e. 11 variables)

R^2 is increased to 94.19% (and Adjusted R^2 to 87.80%) using this method (again not all slopes are significantly different to zero) but I will proceed using all these variables.



Validate Assumptions

Top Left : Can assume normality since graph is approximately linear.

Top Right : Slight problem with assumption of constant variance as points not evenly spread about zero

Regression Analysis: Aroma_1 versus Cd_1, Ni_1, Cu_1, Ba_1, Cr_1, Pb_1, ...

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	11	24.7018	2.2456	14.73	0.000
Cd_1	1	1.4278	1.4278	9.37	0.012
Ni_1	1	1.1467	1.1467	7.52	0.021
Cu_1	1	0.9327	0.9327	6.12	0.033
Ba_1	1	5.3798	5.3798	35.30	0.000
Cr_1	1	5.8275	5.8275	38.24	0.000
Pb_1	1	7.5862	7.5862	49.78	0.000
B_1	1	3.3957	3.3957	22.28	0.001
Na_1	1	2.0367	2.0367	13.36	0.004
Ca_1	1	4.6883	4.6883	30.76	0.000
P_1	1	0.5792	0.5792	3.80	0.080
K_1	1	1.6834	1.6834	11.05	0.008
Error	10	1.5241	0.1524		
Total	21	26.2259			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.390396	94.19%	87.80%	24.58%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.08	1.10	4.60	0.001	
Cd_1	-27.50	8.98	-3.06	0.012	7.63
Ni_1	3.70	1.35	2.74	0.021	2.79
Cu_1	-1.821	0.736	-2.47	0.033	2.10
Ba_1	-9.60	1.62	-5.94	0.000	3.64
Cr_1	-64.5	10.4	-6.18	0.000	3.51
Pb_1	3.748	0.531	7.06	0.000	6.81
B_1	0.3272	0.0693	4.72	0.001	2.37
Na_1	0.01958	0.00536	3.66	0.004	2.54
Ca_1	-0.04449	0.00802	-5.55	0.000	3.63
P_1	-0.00995	0.00510	-1.95	0.080	3.45
K_1	0.00494	0.00149	3.32	0.008	6.67

Regression Equation

Aroma_1 = 5.08 - 27.50 Cd_1 + 3.70 Ni_1 - 1.821 Cu_1 - 9.60 Ba_1 - 64.5 Cr_1
+ 3.748 Pb_1 + 0.3272 B_1 + 0.01958 Na_1 - 0.04449 Ca_1 - 0.00995 P_1
+ 0.00494 K_1

Fits and Diagnostics for Unusual Observations

Obs	Aroma_1	Fit	Resid	Std Resid	
1	3.300	3.078	0.222	2.18	R
7	7.700	7.198	0.502	2.04	R

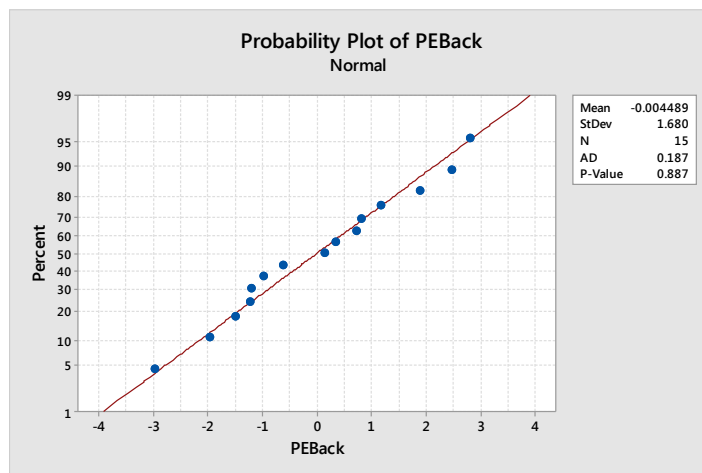
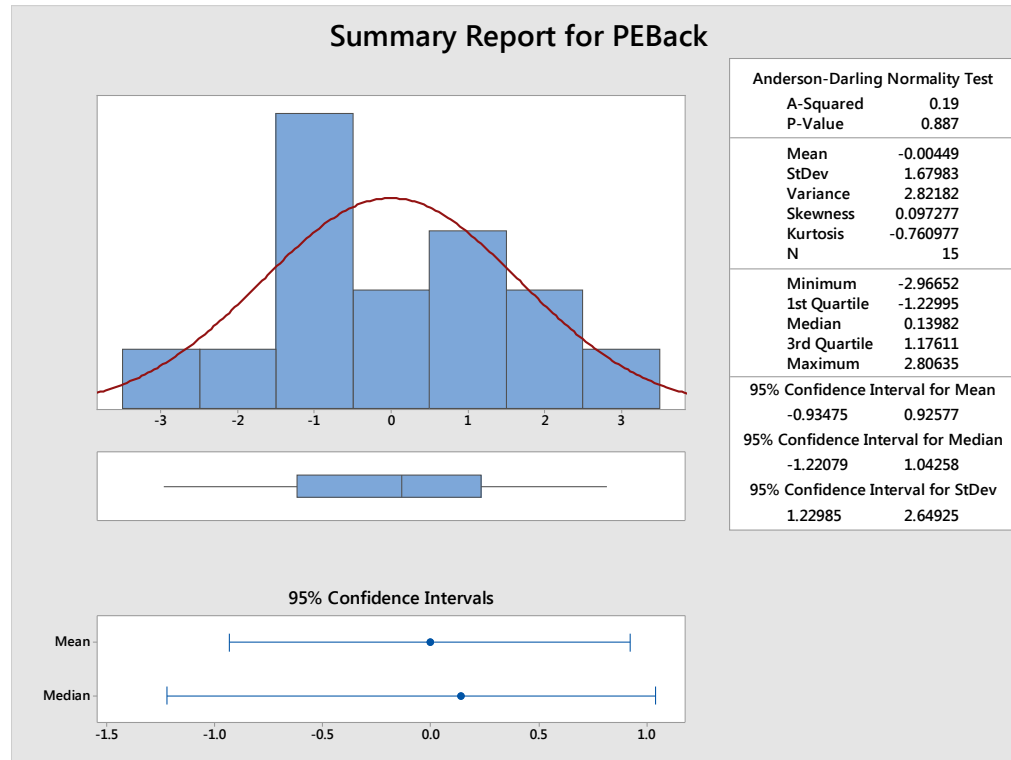
R Large residual

Best Model using Backward Stepping for Training Set

Aroma_1 = 5.08 - 27.50 Cd_1 + 3.70 Ni_1 - 1.821 Cu_1 - 9.60 Ba_1 - 64.5 Cr_1
+ 3.748 Pb_1 + 0.3272 B_1 + 0.01958 Na_1 - 0.04449 Ca_1 - 0.00995 P_1
+ 0.00494 K_1

Predict the Aroma Scores for the Test Set Model is

Predicted Aroma = $5.08 - 27.50 \cdot \text{Cd}_0 + 3.70 \cdot \text{Ni}_0 - 1.821 \cdot \text{Cu}_0 - 9.60 \cdot \text{Ba}_0 - 64.5 \cdot \text{Cr}_0 + 3.748 \cdot \text{Pb}_0 + 0.3272 \cdot \text{B}_0 + 0.01958 \cdot \text{Na}_0 - 0.04449 \cdot \text{Ca}_0 - 0.00995 \cdot \text{P}_0 + 0.00494 \cdot \text{K}_0$



Paired T-Test and CI: Aroma_0, PredictedBack

Paired T for Aroma_0 - PredictedBack

	N	Mean	StDev	SE Mean
Aroma_0	15	4.993	1.024	0.264
PredictedBack	15	4.998	1.631	0.421
Difference	15	-0.004	1.680	0.434

95% CI for mean difference: (-0.935, 0.926)

T-Test of mean difference = 0 (vs ≠ 0): T-Value = -0.01 P-Value = 0.992

Discussion

Prediction Errors are Normally Distributed (Normal probability plot is linear)

Paired t-test ($p=0.992$) shows we cannot reject H_0 : Mean Difference = 0 in favour of H_1 : Mean difference doesn't equal zero at 5% so on average we have a good model

However, in one instance, we have a prediction error of 2.81. So not always very accurate.

FINDING THE BEST MODEL USING BEST SUBSET

Using the same Training and Test Sets.

Note that to make reading the output easier restricted the output to only the Best Model at each stage (see options in the dialogue box)

Best Subsets Regression: Aroma_1 versus Cd_1, Mo_1, ...

Response is Aroma_1

						C	M	M	N	C	A	B	C	S	P	M				S	N	C
						d	o	n	i	u	l	a	r	r	b	B	g	i	a	a	P	K
Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	65.1	63.3	56.0	28.9	0.67663									X								
2	72.5	69.6	62.5	21.0	0.61655							X	X									
3	76.7	72.8	67.7	17.3	0.58321					X	X	X										
4	78.9	74.0	69.2	16.3	0.57005					X	X	X								X		
5	81.0	75.1	66.7	15.5	0.55759		X			X	X	X								X		
6	83.5	77.0	60.4	14.1	0.53630	X	X			X	X	X								X		
7	87.6	81.4	69.5	10.7	0.48244		X		X			X	X						X	X		X
8	89.6	83.2	42.6	9.9	0.45788	X				X		X	X	X	X	X	X	X	X			
9	91.5	85.1	0.0	9.4	0.43071	X				X		X	X	X	X	X	X	X	X		X	
10	92.7	86.0	0.0	9.8	0.41747	X		X		X		X	X	X	X	X	X	X	X		X	
11	94.2	87.8	24.6	9.8	0.39040	X			X	X		X	X	X	X				X	X	X	X
12	95.2	88.7	29.6	10.5	0.37570	X			X	X		X	X	X	X				X	X	X	X
13	95.9	89.2	21.2	11.5	0.36806	X			X	X	X	X	X	X	X	X			X	X	X	X
14	96.7	90.0	30.2	12.5	0.35398	X			X	X	X	X	X	X	X	X			X	X	X	X
15	97.0	89.4	32.2	14.0	0.36315	X	X		X	X	X	X	X	X	X	X			X	X	X	X
16	97.0	87.4	0.0	16.0	0.39639	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X
17	97.0	84.4	0.0	18.0	0.44203	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

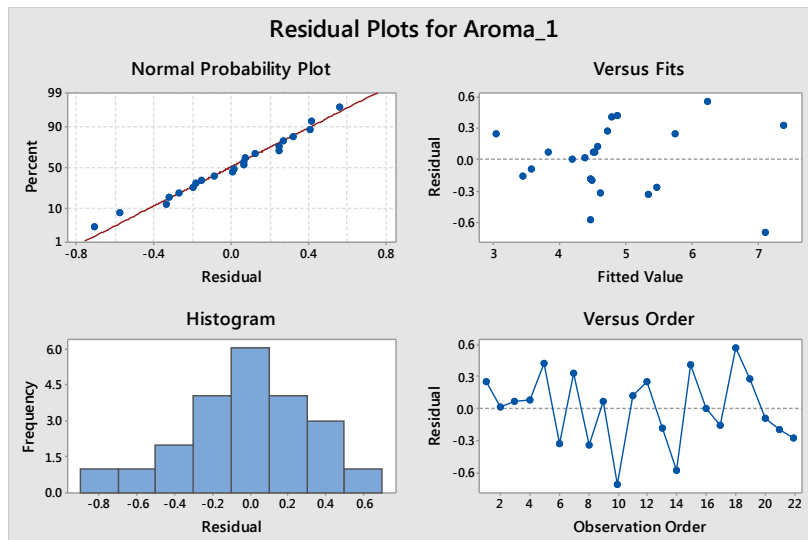
We can either use Maximum Adj R^2 or Minimum Mallows Cp to decide on the Best Model

Maximising Adj R^2 gives model involving Cd_1, Ni_1,Cu_1,Al_1,Ba_1,Cr_1,Sr_1,Pb_1,B_1, Si_1,Na_1,Ca_1,P_1 and K_1 ($R^2=96.7\%$ and adj $R^2= 90.0\%$)

Minimising Mallows Cp gives model involving Cd_1, Cu_1, Cr_1,Pb_1,B_1,Mg_1,Si_1, Na_1 and P_1 ($R^2=91.5\%$ and adj $R^2= 85.1\%$)

I will investigate the second model as it uses fewer variables.

So best model involves : Cd_1, Cu_1, Cr_1,Pb_1,B_1,Mg_1,Si_1, Na_1 and P_1



Validate Assumptions

Top Left : Can assume normality since graph is approximately linear.

Top Right : Slight problem with assumption of constant variance as points not evenly spread about zero

Regression Analysis: Aroma_1 versus Cd_1, Cu_1, Cr_1, Pb_1, B_1, Mg_1, ...

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	9	23.9997	2.6666	14.37	0.000
Cd_1	1	3.6330	3.6330	19.58	0.001
Cu_1	1	3.2725	3.2725	17.64	0.001
Cr_1	1	2.5938	2.5938	13.98	0.003
Pb_1	1	7.2738	7.2738	39.21	0.000
B_1	1	12.2604	12.2604	66.09	0.000
Mg_1	1	6.7635	6.7635	36.46	0.000
Si_1	1	1.4968	1.4968	8.07	0.015
Na_1	1	1.7321	1.7321	9.34	0.010
P_1	1	0.4994	0.4994	2.69	0.127
Error	12	2.2262	0.1855		
Total	21	26.2259			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.430714	91.51%	85.15%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7.51	1.09	6.92	0.000	
Cd_1	-39.44	8.91	-4.43	0.001	6.17
Cu_1	-3.361	0.800	-4.20	0.001	2.04
Cr_1	-31.21	8.35	-3.74	0.003	1.84
Pb_1	3.370	0.538	6.26	0.000	5.75
B_1	0.6049	0.0744	8.13	0.000	2.24
Mg_1	-0.02453	0.00406	-6.04	0.000	1.40
Si_1	-0.0360	0.0127	-2.84	0.015	1.86
Na_1	0.01562	0.00511	3.06	0.010	1.90
P_1	-0.00889	0.00542	-1.64	0.127	3.19

Regression Equation

$$\text{Aroma}_1 = 7.51 - 39.44 \text{ Cd}_1 - 3.361 \text{ Cu}_1 - 31.21 \text{ Cr}_1 + 3.370 \text{ Pb}_1 + 0.6049 \text{ B}_1 - 0.02453 \text{ Mg}_1 - 0.0360 \text{ Si}_1 + 0.01562 \text{ Na}_1 - 0.00889 \text{ P}_1$$

Fits and Diagnostics for Unusual Observations

Obs	Aroma_1	Fit	Resid	Std Resid	
1	3.300	3.054	0.246	2.90	R
10	6.400	7.109	-0.709	-2.27	R

R Large residual

Obs	Unusual	Obs	Unusual	Obs	Unusual	Obs	Unusual
10	0.043	6.4000	7.1089	0.2961	-0.7089	-2.27R	

R denotes an observation with a large standardized residual.

Best Model using Best Subset Regression for the Training Set is

Aroma_1 = 7.51 - 39.44 Cd_1 - 3.361 Cu_1 - 31.21 Cr_1 + 3.370 Pb_1+ 0.6049 B_1 - 0.02453 Mg_1 - 0.0360 Si_1 + 0.01562 Na_1 - 0.00889 P_1

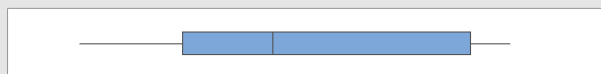
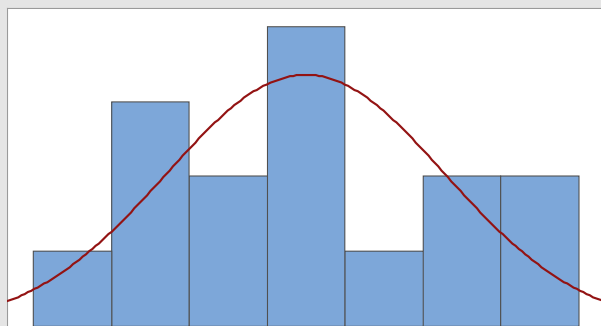
Predict the Aroma Scores for the Test Set

Model is

Predicted Aroma=

$7.51 - 39.44 \cdot \text{Cd}_0 - 3.361 \cdot \text{Cu}_0 - 31.21 \cdot \text{Cr}_0 + 3.370 \cdot \text{Pb}_0 + 0.6049 \cdot \text{B}_0 - 0.02453 \cdot \text{Mg}_0 - 0.0360 \cdot \text{Si}_0 + 0.01562 \cdot \text{Na}_0 - 0.00889 \cdot \text{P}_0$

Summary Report for PEBest



Anderson-Darling Normality Test

A-Squared	0.41
P-Value	0.295

Mean	-0.00563
StDev	1.78287
Variance	3.17863
Skewness	0.25440
Kurtosis	-1.09893
N	15

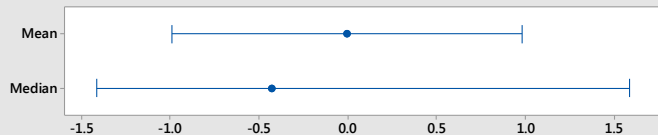
Minimum	-2.90269
1st Quartile	-1.59049
Median	-0.42710
3rd Quartile	2.10952
Maximum	2.62471

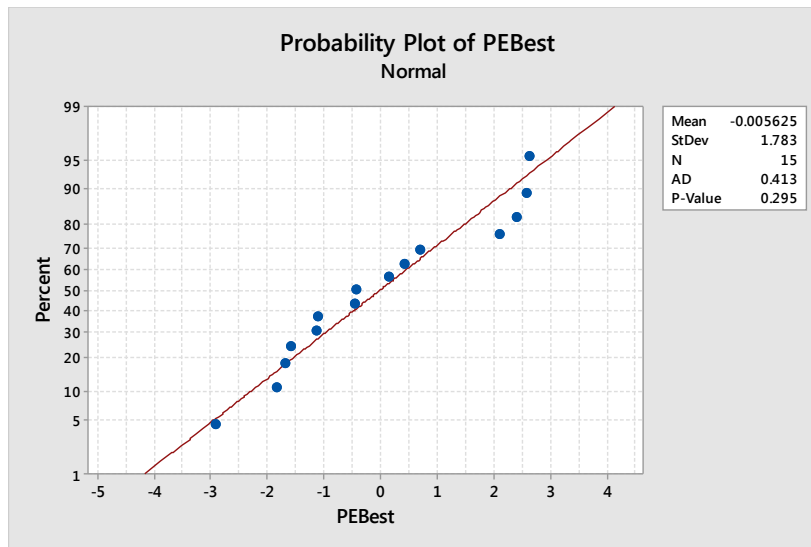
95% Confidence Interval for Mean	-0.99295	0.98170
----------------------------------	----------	---------

95% Confidence Interval for Median	-1.41501	1.58819
------------------------------------	----------	---------

95% Confidence Interval for StDev	1.30529	2.81176
-----------------------------------	---------	---------

95% Confidence Intervals





Discussion

Prediction Errors are Normally Distributed
(Normal probability plot is linear)

Paired t-test ($p=0.990$) shows we cannot reject H_0 : Mean Difference = 0 in favour of H_1 : Mean difference doesn't equal zero at 5% so on average we have a good model

However, in one instance, we have a prediction error of 2.62. So not always very accurate.

Paired T-Test and CI: Aroma_0, PredictedBest

Paired T for Aroma_0 - PredictedBest

	N	Mean	StDev	SE Mean
Aroma_0	15	4.993	1.024	0.264
PredictedBest	15	4.999	1.945	0.502
Difference	15	-0.006	1.783	0.460

95% CI for mean difference: (-0.993, 0.982)

T-Test of mean difference = 0 (vs \neq 0): T-Value = -0.01 P-Value = 0.990

We have now developed and tested models using 4 equally valid techniques.

The 'Best' model for Aroma Scores depends on how accurate your predictions need to be. We could not reject that the average prediction error equalled zero for any of the models when used on the Test Set. However, in particular cases the predictions were inaccurate!!

So it is up to the investigator to look at the mean and sd for the prediction errors and the cost of taking measurements!!

Stepwise Regression

Aroma=6.869 - 1.389 Cu + 6.23 Ba - 3.936 Sr

Forward Stepping

Aroma=7.499 - 11.56 Cd + 4.50 Mo - 1.757 Cu + 5.70 Ba - 3.420 Sr - 0.01274 Ca

Backward Stepping

**Aroma= 5.08 - 27.50 Cd + 3.70 Ni - 1.821 Cu - 9.60 Ba - 64.5 Cr
+ 3.748 Pb + 0.3272 B + 0.01958 Na - 0.04449 Ca - 0.00995 P
+ 0.00494 K**

Best Subset

**Aroma= 7.51 - 39.44 Cd - 3.361 Cu - 31.21 Cr + 3.370 Pb + 0.6049 B
- 0.02453 Mg - 0.0360 Si + 0.01562 Na - 0.00889 P**

Observed and Predicted for all four methods for the TEST SET

Row	Aroma_0	Predicted	PredForward	PredictedBack	PredictedBest
1	3.9	3.40767	3.07635	2.00489	1.79048
2	5.6	4.71788	6.13933	6.82995	8.50269
3	4.8	3.60547	4.60834	7.76652	6.63588
4	4.3	3.58694	3.88082	5.79142	5.97207
5	5.1	5.36417	5.13550	6.30544	6.22067
6	3.3	0.21840	1.06252	2.48140	0.88760
7	5.9	5.37815	4.91297	4.72389	5.46855
8	7.1	6.34506	6.49429	4.29365	4.47529
9	5.5	5.87611	5.66002	3.02041	2.92501
10	6.3	6.52806	6.31755	6.16018	5.58624
11	5.5	6.34033	6.46327	4.76577	5.92710
12	4.1	3.59191	3.73822	4.71811	4.54504
13	3.9	4.33890	4.59794	5.85505	5.49049
14	5.1	5.24206	5.09942	6.08561	6.21117
15	4.5	4.24558	4.47697	4.16504	4.34608