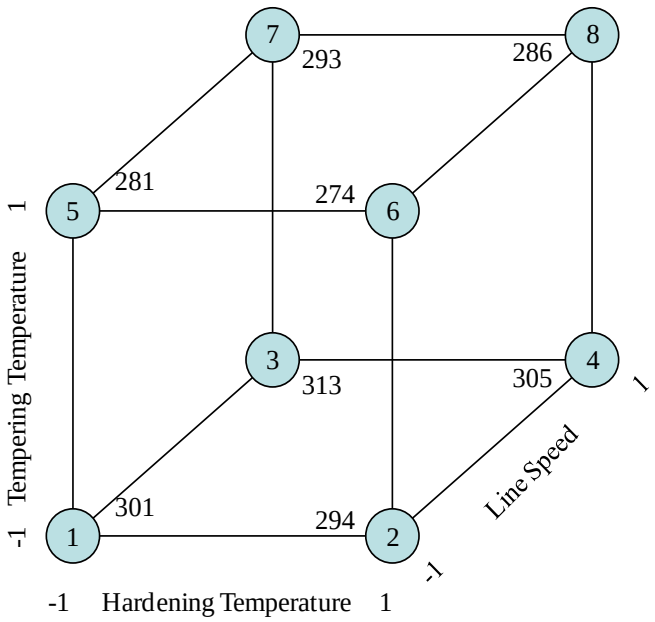
DOE/Multiple Linear Regression Instructions for SPC for MS Excel Version 4



Thank you for selecting our software package. This manual contains the DOE and multiple linear regression instructions for SPC for MS Excel Version 4. This program is owned by BPI Consulting, LLC. This program cannot be copied or used unless under license with BPI Consulting, LLC. BPI Consulting, LLC is not liable for any decisions made based on the use of this software package.

Requirements: This program is a Microsoft Excel® add-in. You must Microsoft Excel® for this program to work. This program supports any version of Excel from 2000 on.

BPI Consulting, LLC
20314 Lakeland Falls
Cypress, TX 77433
800-274-2874
281-304-9504
www.spcforexcel.com
mailto:support@spcforexcel.com

Table of Contents

Introduction to DOE....4
Introduction to Multiple Linear Regression....5
References....6
Setting Up the Two Level Experimental Design....7
Analysis of the Two Level Experimental Design....10
Two Level Experimental Design Output: Calculations and Interpretation....10
- All Factors Analysis Worksheet....11
  - Design Table....11
  - Range Control Chart: Checking for Statistical Control and Estimating Sigma....12
  - ANOVA Table for Factors and Interactions....13
  - ANOVA Table for the Model....14
  - Factor Information....15
  - Model Containing All Factors....15
- Normal Plot of Effects Sheet....16
- Half-Normal Plot of Effects Sheet....16
- Effect Graphs Worksheet....17
- Two Factor Plots....17
- All Factors Residual Info Worksheet....18
- Residual Plots Sheet....19
- DOE Optimization....20

Adding/Removing Terms in the Model for Two Level Experimental Designs....22
Setting Up the One and Two Factor Analysis of Variance....24
Analysis of Variance Output: Calculations and Interpretation....25
- ANOVA Worksheet....26
  - Table of Results....26
  - ANOVA Table....26
  - ANOVA Table for the Model....27
  - Treatment Mean Confidence Intervals....28
  - Residuals and Residual Plots Sheets....28
  - Scatter Sheet....28

Setting Up the Multiple Linear Regression....29
Multiple Linear Regression Output....30
- Original Regression Summary Worksheet....31
  - ANOVA Table....31
  - Coefficients....32

Regression Statistics 32
Original Residuals Data Worksheet 33
Residual Plots Sheet 34
Revising the Regression 36
Remove Variables 36
Remove Observations 36
Transform Y Variable 36

DOE Instructions for SPC/DOE for MS Excel

Introduction to DOE

The experimental design module for SPC for MS Excel contains the following experimental designs:

- One factor analysis of variance
- Two factor analysis of variance
- Two level full factorial designs (up to 7 factors)
- Two level fractional factorial designs (29 designs to choose from for up to 15 factors)
- Two level Plackett-Burman designs (up to 27 factors)

The two level designs include the following:

- Design table analysis using Yates' algorithm
- ANOVA table for the factors and interactions
- ANOVA table for the model
- Average, standard deviation, coefficient of variation, R^2, adjusted R^2, PRESS and R^2 prediction
- Factor coefficients with 95% confidence limits
- Model using coded factors
- Model using actual factors
- Normal probability plot of the effects
- Half-normal plot of the effects where effects to be included in the model can be selected and the analysis re-ran with the new effects
- Plot of main effects, two factor effects and three or more factor effects
- Two factor plots
- Range control chart when replicates are run to check for out of control situations
- Curvature check when center points are run
- Residual analysis
    - Raw residuals
    - Leverage
    - Standardized residuals
    - Internally studentized residuals
    - Externally studentized residuals
    - DFFITS
    - Cook's distance
- Residual plots for each type of residual
    - Normal plot of residuals
    - Residuals versus predicted results
    - Residuals versus actual run number
- Other plots
    - Leverage versus actual run number
    - DFFITs versus actual run number
    - Cook's distance versus actual run number
    - Predicted values versus predicted values
- Optimization
    - Two factor plots

The two level designs can handle replications, center points and multiple responses. If replications are run, there is the option to analyze the range results to determine what factors and/or interactions affect variability.

The one factor and two factor ANOVA contains the following outputs:

- ANOVA table for the factor
- ANOVA table for the model
- Scatter diagram of results
- Treatment means with 95% confidence limits
- Comparison of Means:
    - Fisher Least Significant Difference Method
    - Bonferroni's Method
    - Tukey's Method
- Test for equality of variances:
    - Bartlett's Test
    - Levene's Method
- Residual analysis (same as in the two-level experimental designs)

Introduction to Multiple Linear Regression

The multiple linear regression module of this software package allows you to determine the relationship between several independent or predictor variables and a dependent or response variable. The multiple linear regression analysis provides the following output:

- ANOVA table for the model
- Coefficients table
    - Coefficient
    - Standard Error
    - t statistic
    - 95% upper and lower confidence interval for the coefficient
    - VIF (variation inflation factor)
    - Standardized coefficients
- Regression statistics
    - R
    - R squared
    - Adjusted R squared
    - Mean
    - Standard error
    - Coefficient of variation
    - Observations
    - Durbin-Watson statistic
    - PRESS
    - R squared prediction
- Residual Analysis (same as in the experimental design section)

This technique also allows you to easily remove variables and observations as well as transform the response variable using one of the following: square root, arcsine, reciprocal square root, reciprocal or Box-Cox transformation.

References

The software is built primarily around the information contained in the following publications:

1. Montgomery, D. C., Peck, E., and Vining, G. G, Introduction to Linear Regression Analysis, 4th Edition, John Wiley & Sons, 2006
2. Montgomery, D.C., Design and Analysis of Experiments, 6th Edition, John Wiley & Sons, 2005
3. Box, G. E. P., Hunter, W. G., and Hunter, J. S., Statistics for Experimenters, John Wiley & Sons, 1978
4. Experimental Design Manual, Vista Chemical Company, 1986

References 1-3 can be purchased from many on-line bookstores and are highly recommended for those who want to learn more about experimental design techniques and multiple linear regression.

Setting Up the Two Level Experimental Design

This procedure applies to the full factorial designs, the fractional factorial designs, and the Plackett-Burman designs. There are some slight differences between the three.

The experimental design program is run from the DOE icon on the SPC toolbar. It is often advantageous to have the responses and the factors with their high and low levels already entered into a worksheet. The program gives you the option of entering this information into dialog boxes and from ranges on a worksheet. The ranges on a worksheet work best.

To understand how the design is set up, we will use an example from Montgomery's book on experimental design. The example involves a 2^3 full factorial design that is used to develop a nitride etch process on a single-wafer plasma etching tool. There are three design factors:

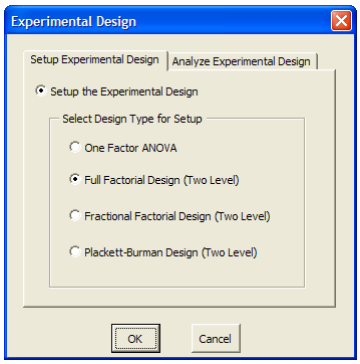
1. Gap between the electrodes in centimeters (0.8, 1.2)
2. C_2F_6 gas flow in SCCM (125, 200)
3. RF power applied to the cathode in watts (275, 325)

Each factor was run at two levels shown above and the experiment was replicated twice. The response variable is the etch rate for silicon nitride.

To set up this design, the following information was entered into an Excel worksheet:

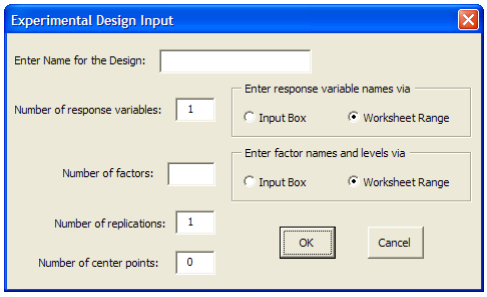
	A	B	C
1	Nitride Etch		
2			
3	Factor Information		
4	Name	Low Level	High Level
5	Gap	0.8	1.20
6	C2F6 Flow	125	200
7	Power	275	325
8			
9			

This makes it easier to enter the required information to setup the design. Select the DOE () icon from the SPC menu. The following dialog box will appear.



Select the experimental design option you want. In this case, it is a full factorial design (two levels). Then select OK.

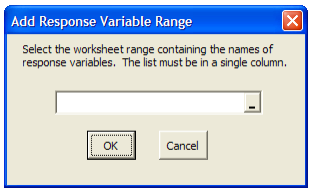
You will see the dialog box to the right which requires you to enter the following:



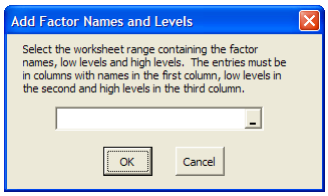
- Name for the design (e.g., Etch Rate)
- Number of response variables (the default value is 1)
- Enter the response variable names via an input box or worksheet range (worksheet range is the default)
- Number of factors (in this example, 3)
- Enter factor names and levels via an input box or worksheet range (worksheet range is the default)
- Number of replications (in this example, 2)
- Number of center points (in this example, 0)

It should be noted that if center points are added, they are used to determine if curvature is present and will be used to estimate the variability (error) if the number of replications is 1. For the replications > 1, the replicated runs are used to estimate the variability (error).

Once you select OK, you will see the two dialog boxes below in the order given if you selected the worksheet option for entering the information.

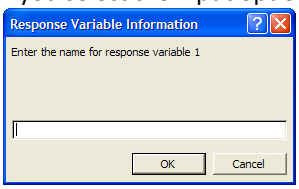


Enter the worksheet range containing the name of the response variable (cells A1 in the example above)

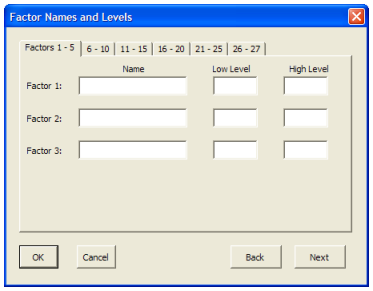


Enter the worksheet range containing the factor name, low level and high level (cells A5 to C7 above). Note: the factor levels must be numeric. If you have discrete levels, such as day and night, assign numeric values to them (such as day = 1 and night = 1).

If you select the input options, you will see the two dialog boxes below in the order given.



Enter the name of the response. You will get a separate dialog box for each response variable.



Enter the factor name, low level and high level for each factor.

Once the above information is entered, the design is developed in a new worksheet as shown below.

	A	B	C	D	E	F	G
2	**Type:**	3 Factor Full Factorial					
3							
4	**Response Variables**						
5	Nitride Etch						
6							
7	**Factor Information**						
8	Factor	Name	Low Level	High Level			
9	A	Gap	0.8	1.20			
10	B	C2F6 Flow	125	200			
11	C	Power	275	325			
12							
13							
14	Actual Run Order	Standard Run Order	A	B	C	Nitride Etch	
15	1	7	0.8	200	325	1075	
16	2	3	0.8	200	275	633	
17	3	4	1.2	200	275	642	
18	4	5	0.8	125	325	1037	
19	5	7	0.8	200	325	1063	
20	6	4	1.2	200	275	635	
21	7	2	1.2	125	275	650	
22	8	3	0.8	200	275	601	
23	9	1	0.8	125	275	550	
24	10	2	1.2	125	275	669	
25	11	8	1.2	200	325	729	
26	12	6	1.2	125	325	749	
27	13	6	1.2	125	325	868	
28	14	1	0.8	125	275	604	
29	15	5	0.8	125	325	1052	
30	16	8	1.2	200	325	860	
31							

The new worksheet contains the name of the design, the type of design, the response variable, and the factor information. It also includes the experimental runs to be completed in a randomized order.

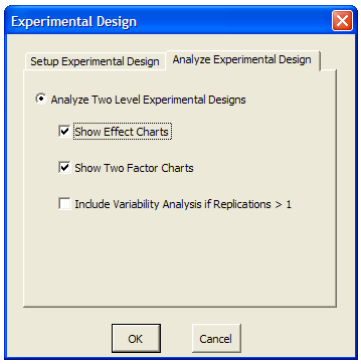
The actual run number is the order the experiment should be carried in. The standard run number represents the standard runs in an experimental design table.

The results for each experimental run are then entered into the design table under the response variable (in this example, nitride etch). This has been done already in the worksheet shown here.

You are now ready to analyze the experimental results.

Analysis of the Two Level Experimental Design

Once the experimental design has been run and the results entered into the worksheet as shown above, you are ready to analyze the results. This process is started by selecting the DOE icon on the SPC for Excel menu. You should be on the worksheet containing the experimental results. You will see the dialog below.



Select from the following options:

- Show Effect Charts: these charts contrast the high and low level effects for each factor and interaction
- Show Two Factor Charts: these charts show the relationship between each pair of factors
- Include Variability Analysis if Replications > 1: this option will generate a separate analysis of the range results to determine what factors/interactions impact variability

Once you select OK, the program will perform the analysis. A new workbook is created and the data is transferred into the workbook. All the calculations and output is done in the new workbook. The output from the analysis is described below.

Two Level Experimental Design Output: Calculations and Interpretation

The new workbook contains the following sheets initially:

- DOE Data: this worksheet contains the data that was transferred over
- All Factor Analysis: this worksheet contains the data for all the factors/interactions including:
  - Design table analyzed using Yates' Algorithm
  - Range chart results if replications were run
  - ANOVA table based on all factors and interactions
  - ANOVA for the model (containing all the factors)
  - Average, standard deviation, coefficient of variation, R^2, adjusted R^2, PRESS and R^2 prediction
  - Factor information include coefficient, degrees of freedom, standard error, and 95% confidence limits
  - Model containing all factors based on coded and actual levels
- Normal Probability Plot of Effects
- Half-Normal Plot of Effects
- Effect Charts (if that option was selected)
- Two Factor Plots (if that option was selected)
- Residuals Plots: contains the normal probability plot of the raw residuals initially but has options for many more
- All Factors Residuals Info: contains the following:
  - Standard run number
  - Actual run number
  - Observed value
  - Predicted value
  - Raw residuals
  - Leverage
  - Standardized residuals
  - Internally studentized residuals
  - Externally studentized residuals
  - DFFITS
  - Cook's distance

Details on each worksheet (except for DOE Data worksheet which contains the raw data) are given below for multiple replications. The output for a single replication is very similar.

All Factors Analysis Worksheet

This worksheet contains the results for a model containing all the factors and interactions in the experimental design.

Design Table

It begins with the classic design table which shows the standard runs and the results. The design table for the etch rate example is shown below. The significant effects are in bold in the row labeled "Effects."

Design Table												
Standard Run Order	Mean	*A*	B	*C*	AB	*AC*	BC	ABC	Average	Range		
1	+	-	-	-	+	+	+	-	577	54		
2	+	+	-	-	-	-	+	+	659.5	19		
3	+	-	+	-	-	+	-	+	617	32		
4	+	+	+	-	+	-	-	-	638.5	7		
5	+	-	-	+	+	-	-	+	1044.5	15		
6	+	+	-	+	-	+	-	-	808.5	119		
7	+	-	+	+	-	-	+	-	1069	12		
8	+	+	+	+	+	+	+	+	794.5	131		
Sum +	6208.5	2901	3119	3716.5	3054.5	2797	3100	3115.5				
Sum -	0	3307.5	3089.5	2492	3154	3411.5	3108.5	3093				
Overall	6208.5	6208.5	6208.5	6208.5	6208.5	6208.5	6208.5	6208.5				
Difference	6208.5	-406.5	29.5	1224.5	-99.5	-614.5	-8.5	22.5				
Effect	776.063	**-101.625**	7.375	**306.125**	-24.875	**-153.625**	-2.125	5.625				
SS		41310.563	217.563	374850.063	2475.063	94402.563	18.063	126.563				
MSE	49.703											

The significant effects are in bold in the effects row. These are larger than MSE.

The standard run number is given in Column A, followed by columns for the mean and each factor and interaction in the design. The last two columns are the average and range. The average is the average result for the runs at each standard condition. For example, standard run 1 was run twice during the design (actual runs 9 and 14). The two results were 550 and 604. So, the average is (550+604)/2 = 557. The range is the maximum – minimum for the standard run. For standard run 1, the range is 604 – 550 = 54.

The information below the design table starting with the row labeled "Sum +" is the table analysis using Yates's algorithm. This will be demonstrated using Factor A results. The rows perform the following functions:

- Sum +: sums the results for a factor at its high level (+);
  - Factor A: Sum + = 659.5 +638.5 +808.5 + 794.5 = 2901
- Sum -: sums the results for a factor at its low level (-)
  - Factor A: Sum - = 577 + 617 + 1044.5 + 1069 = 3307.5
- Overall: The sum of the sum + and sum – values (this was used as a check on calculations and is the same for all factors)
  - Factor A: Overall = (Sum +) + (Sum -) = 2901 + 3307.5 = 6208.5
- Difference: The difference between the sum + and the sum – values; this represents the difference between the sum of the results at the factor's high level and the sum of the results at the factor's low level.
  - Factor A: Difference = (Sum +) - (Sum -) = 2901 - 3307.5 = -406.5
- Effect: This is the effect of the factor. It is determined by dividing the difference by the number of plus signs in the column. This effect is the difference in the average for the results at the high level of the

factor and the average of the results at the low level of the factor. (Note: the effect under the mean column is the average of all the factorial runs.)
    - Factor A: Effect = Difference/4 = -406.5/4 = -101.625
- SS: This is the sum of squares for the factor. This the number of replications times the difference squared divided by the total number of factorial runs (N).
    - Factor A: SS = (NReps*Difference)^2/N = (2*-406.5)^2/16 = 41,310.56

The last row starts in this section is labeled MSE. This stands for Minimum Significant Effect. It is one way of determining which effects are significant. The equation for MSE is:

$$\text{MSE} = t_{(0.05,\upsilon)}\sigma'\sqrt{4/N}$$

where $t_{(0.05,\upsilon)}$ is the t value for 95% confidence, υ is the degrees of freedom (number of factorial observations – number of cells), σ' is the estimated standard deviation obtained from the range values (see below), and N is the total number of factorial runs. In this example, the following can be calculated for MSE:

υ = Number of factorial observations – number of cells = 16 – 8
$t_{(0.05,\upsilon)}$ = 2.306
σ' = $\bar{R}/d_2$ = 48.625/1.128 = 43.10727 where $\bar{R}$ is the average range and d_2 is control chart constant that depends on subgroup size (the number of replications)
N = 16
MSE = $t_{(0.05,\upsilon)}\sigma'\sqrt{4/N}$ = $2.306(43.10727)\sqrt{4/16}$ = 47.7028

This value of MSE is compared to the effects in the design table. Any absolute value of an effect that is greater than MSE is considered significant. The program changes these effects to bold (in the effects row). As can be seen from the design table above, A, C and AC are significant effects.

Range Control Chart: Checking for Statistical Control and Estimating Sigma

When multiple replications are run, the program uses the range values to estimate the variability in the process and to check for out of control situations. The output for the range control chart on the All Factors Analysis Worksheet is shown below as well as the calculations.

Range Chart Results		
Rbar	48.625	***The ranges are in statistical control.***
UCLr	158.858	
LCLr	None	

The average range is calculated along with the upper control limit (UCLr) and the lower control limit (LCLr). The range values are compared to the control limits. If none of the range values are beyond these limits, the ranges are in statistical control. If any range is beyond these limits, there is evidence of a special cause of variation that may make the results suspect.

The average range, $\bar{R}$, and the control limits are calculated using the following equations:

$$\bar{R} = \frac{\Sigma \mathrm{R_i}}{k}$$

$$UCLr = D_4\bar{R}$$

$$LCLr = D_3\bar{R}$$

where R_i is the range of standard run i, k is the number of range values, and D_4 and D_3 are control chart constants that depend on the number of replications (the subgroup size).

In this example, k = 8 and the number of replications is 2. The average range and control limits are given by:

$$\bar{R} = \frac{\Sigma R_i}{k} = \frac{389}{8} = 48.625$$

$$UCLr = D_4\bar{R} = 3.267 * 48.625 = 158.8579$$

There is no lower control limit on a range for 2 replications. The values for D_4 and D_3 for various subgroup sizes are available in many publications and on our website. Since no range value is above 158.8579, we conclude that the ranges are in statistical control and there were no special causes present when the experimental design was run. The residuals analysis (see below) will also check to determine if there were any issues present when the design was run. Note that the average range is used to estimate the value of the standard deviation used in the MSE equation above.

ANOVA Table for Factors and Interactions

The next portion of the All Analysis Factor worksheet is the ANOVA table for the factors and interactions. The output for this example is shown below. The significant effects are those with a p-value ≤ 0.05. A p-value is in bold if it is less than 0.05. If the p-value is between 0.05 and 0.20, the p-value is in italics. This effect may or may not be significant. It is border-line and probably should be considered for inclusion in the model. In the example, A, C and AC are significant. This agrees with the result from the design table.

ANOVA Table Based on All Factors and Interactions							
Source	SS	df	MS	F	p-value	% Cont	
A	41310.563	1	41310.563	18.339	0.0027	7.77%	
B	217.563	1	217.563	0.097	0.7639	0.04%	
C	374850.063	1	374850.063	166.411	0.0000	70.54%	***The signficant factors are in red (p <= 0.05). Factors in blue (0.05 < p <=0.20) may or may not be significant.***
AB	2475.063	1	2475.063	1.099	0.3252	0.47%	
AC	94402.563	1	94402.563	41.909	0.0002	17.76%	
BC	18.063	1	18.063	0.008	0.9308	0.00%	
ABC	126.563	1	126.563	0.056	0.8186	0.02%	
Error	18020.5	8	2252.5625			3.39%	
Total	531420.938	15				100.00%	

The columns in the ANOVA table are:

- Source: the source of variation which includes the factors and interactions in the model as well as the error and the total
- SS: sum of squares for each source of variation
  - The sum of squares for the factors and interactions are given in the design table
    - The model sum of squares (SS_{Model}) is the sum of the factors' and interactions' sum of squares
  - The total sum of square (SS_{Total}) is given by the equation below where y_i represents an experimental result and N is the total number of experimental runs
    - $SS_{Total} = \Sigma y_i^2 - (\Sigma y_i)^2/N$
  - The error sum of squares is determined by subtracting the factor and interaction sum of squares from the total sum of squares.
    - $SS_{Error} = SS_{Total} - SS_{Model}$

- df: degrees of freedom
  - The degrees of freedom for each factor and interaction is 1 since these are two level designs
  - The model degrees of freedom (df_{Model}) is equal to the number of factors and interactions in the model
  - The total degrees of freedom (df_{Total}) is equal to the total number of runs minus 1
    - $df_{Total} = N - 1$
  - The error degrees of freed (df_{Error}) is the given by:
    - $df_{Error} = df_{Total} - df_{Model}$
- MS: mean square
  - The mean square for a source is the variance associated with that source and is determined by dividing the source sum of squares by the degrees of freedom for that source.
    - MS = SS/df
- F: value from the F distribution
  - The F Value for a source of variation is used to compare the variance associated with that source with the error variance.
    - F = MS/MSE where MS is the mean square for a source and MSE the mean square error
- p-Value: the probability value that is associated with the F Value for a source of variation
  - It represents the probability of getting a given F Value if the source does not have an effect on the response.
  - If the p-value is <=0.05, it is considered to have a significant effect on the response.
  - A p-value above 0.20 is not considered to have an effect on the response
  - If the p-value is between 0.05 and 0.20, it or may not have a significant effect.
- % Cont: the % of the total sum of squares the source of variation accounts for
  - Smaller p-values will generate larger % contribution
    - % Cont = SS/SS_{Total} where SS is the sum of squares for a given source of variation

ANOVA Table for the Model

The model's ANOVA table is listed next on the All Factors Analysis worksheet. The output for this example is shown below.

ANOVA for Model					
Source	SS	df	MS	F	P-value
Model	513400.438	7	73342.920	32.560	0.0000

Average	776.0625
Standard Deviation	47.461
Coefficient of Variation	6.116
R Squared	96.61%
Adjusted R Squared	93.64%
PRESS	72082
R Squared Prediction	86.44%

The columns in the ANOVA table have been explained above. There are several pieces of information below the ANOVA table. These are:

- *Average:* the average of the experimental runs
- Standard deviation: the square root of the mean square error
- *Coefficient of variation:* the error expressed as a % of the average, 100(Average/Standard Deviation)
- *R Squared:* measures the proportion of the total variability measured explained by the model
  - $R^2 = 1 - \frac{SS_{Error}}{SS_{Model} + SS_{Error}}$
- *Adjusted R Squared:* the value of R^2 adjusted for the size of the model (the number of factors in the model)
  - $R^2_{Adj} = 1 - \frac{SS_{Error}\ /df_{Error}}{\left(\frac{SS_{Model}\ +SS_{Error}}{df_{Model}\ +df_{Error}}\right)}$
- *PRESS:* Predicted Error Sum of Squares is a measure of how well the model will predict new values and is given below where e_i is the i^{th} residual and h_{ii} is the diagonal element of the hat matrix ($H=X(X'X)^{-1}X'$)
  - $PRESS = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$
- *R Squared Prediction*: indication of the predictive capability of the model; the percent of the variability the model would be expected to explain with new data

- $R^2 = 1 - \frac{PRESS}{SS_{Total}}$

Factor Information

Information on the factors is listed next on the All Factors Analysis worksheet. This information includes the factor, degrees of freedom, the standard error and 95% confidence limits. The output for this example is shown to the right. The columns are explained below.

Factor	Coeff	Degrees of Freedom	Standard Error	95% Lower	95% Upper
Intercept	776.063	1	11.865	748.701	803.424
A	-50.813	1	11.865	-78.174	-23.451
B	3.688	1	11.865	-23.674	31.049
C	153.063	1	11.865	125.701	180.424
AB	-12.438	1	11.865	-39.799	14.924
AC	-76.813	1	11.865	-104.174	-49.451
BC	-1.063	1	11.865	-28.424	26.299
ABC	2.813	1	11.865	-24.549	30.174

- *Factor:* the intercept, factors and interactions included in the model
- *Coeff:* the regression coefficients (β) for the factors
- *Degrees of Freedom:* the degrees of freedom associated with the factor (always 1 for two level designs)
- *Standard Error:* the estimated variance of the factor which is defined as the following for n = number of replications and k = the number of factors:
  - se = $\sqrt{\frac{MSE}{n2^k}}$
- *95% Upper and Lower Confidence Limits:* The upper and lower 95% confidence limit around the coefficient; if it contains zero, the factor is usually not significant
- 95% Lower Confidence Limits = $\beta - t_{(0.05,df_{Error})}(se)$
- 95% Upper Confidence Limits = $\beta + t_{(0.05,df_{Error})}(se)$

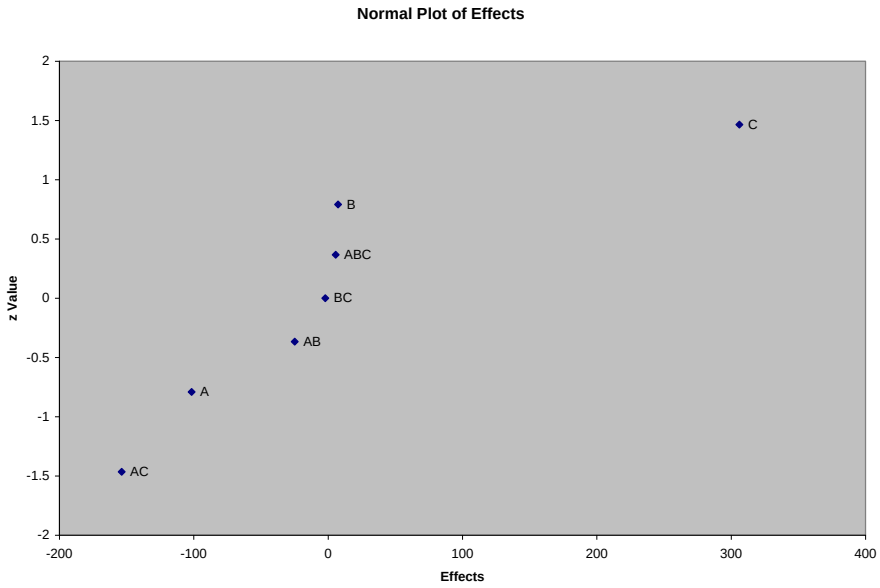
Model Containing All Factors

The last part of the All Factors Analysis worksheet contains the model for all the factors based on the coefficients given above. The coded model is based on the coded levels (-1 to 1) for the factors and interactions. The uncoded model is based on the actual examples. The output for the example is shown below.

Model Containing All Factors			
Coded			
Y=	776.063	Intercept	
+	-50.813	A	
+	3.688	B	
+	153.063	C	
+	-12.438	AB	
+	-76.813	AC	
+	-1.063	BC	
+	2.813	ABC	
UnCoded			
Y=	-6487.333	Intercept	
+	5355.417	Gap	
+	6.597	C2F6 Flow	
+	24.107	Power	
+	-6.158	Gap*C2F6 Flow	
+	-17.800	Gap*Power	
+	-0.016	C2F6 Flow*Power	
+	0.015	Gap*C2F6 Flow*Power	

Normal Plot of Effects Sheet

This sheet contains the normal plot of the effects. The output from this example is shown below. The effects are plotted on the x-axis and the z-values on the y-axis. You can use this chart to determine what effects are significant. Significant effects are those that tend to fall off an imaginary straight line drawn through most of the points. In this case, a straight line fits easily through effects B, ABC, BC and AB. This leaves A, C and AC off the straight line making them probable significant effects. The half-normal plot discussed below represents a better process for doing this.



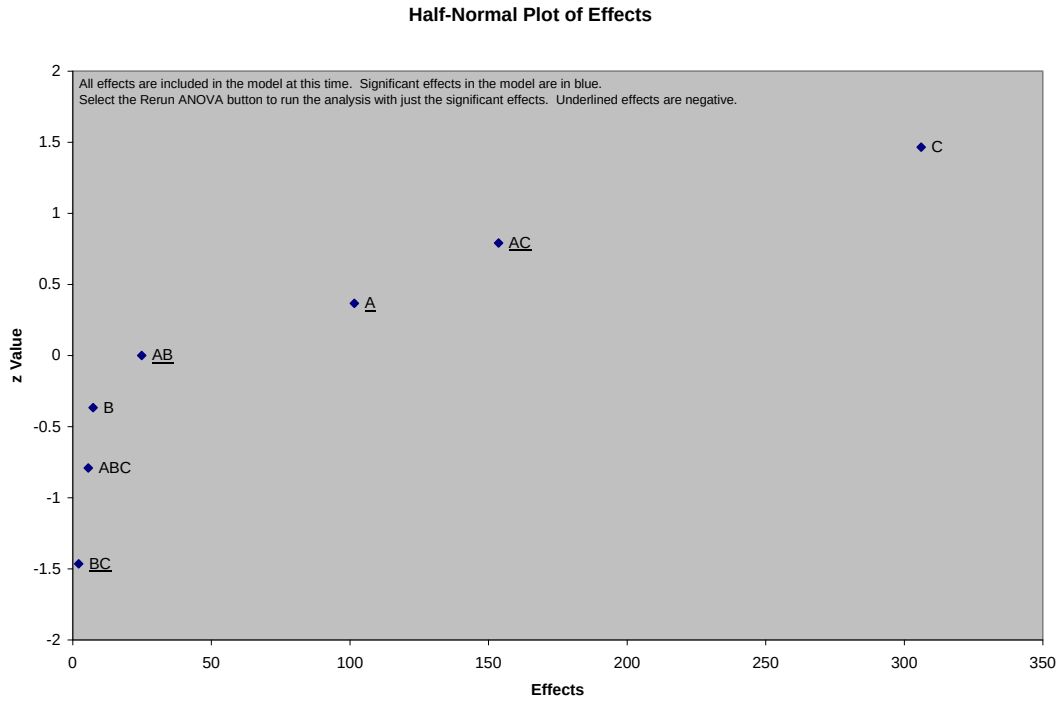
Half-Normal Plot of Effects Sheet

This sheet contains the half-normal plot of effects as shown below. The absolute value of the effects is used in this plot. You will see the following in the upper left-hand corner of this chart the first time it is made:

> "All effects are included in the model at this time. Significant effects in the model are in blue."

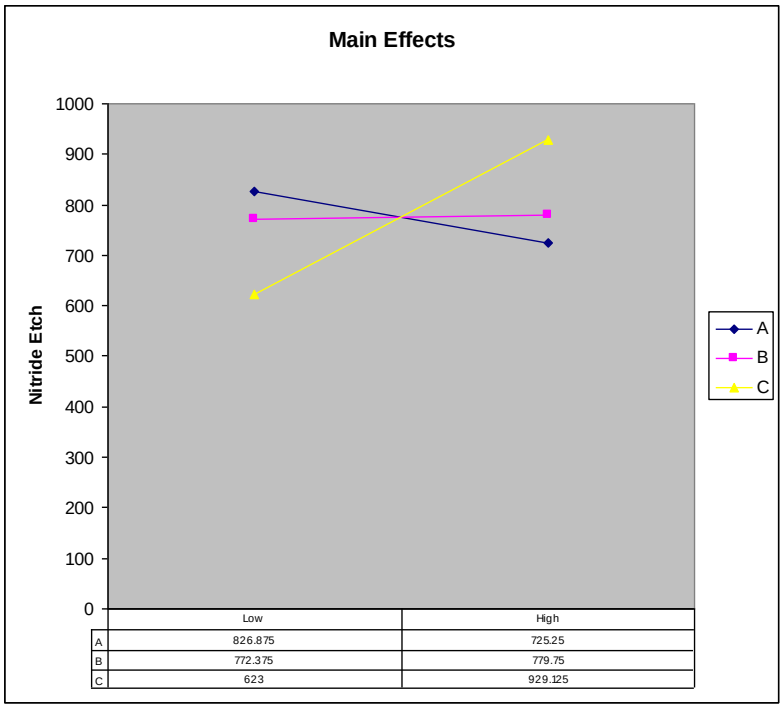
Select the Rerun ANOVA button to run the analysis with just the significant effects. Underlined effects are negative."

All the effects are included in the model when it is first run. On this sheet, the significant effects from the design table analysis are in blue. You can select additional effects to include in the model by selecting the "Select Points" button or you can re-run the ANOVA analysis using the selected effects (in blue) by selecting the "ANOVA" button. This is discussed in more detail below in the "Adding or Removing Effects from the Model."



Effect Graphs Worksheet

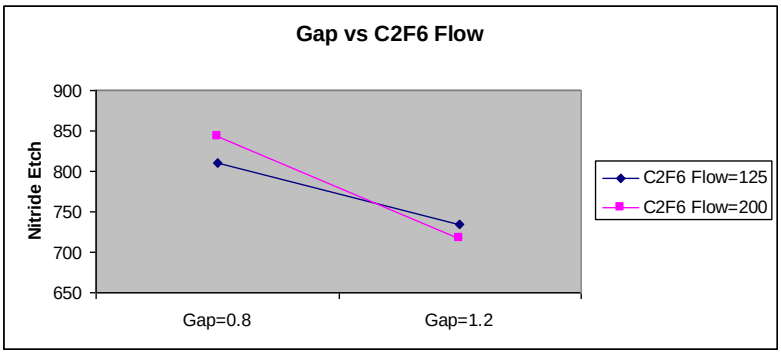
This worksheet contains the effect graphs for the main effects, two factor effects and three or more effects. An example of the effect graphs for the main effects in this model is given below.



This chart plots the average of the runs at the high level of the factor and the average of the runs at the low level of the factor. The steeper the line, the more likely it is that the effect of the factor is significant.

Two Factor Plots

This worksheet contains the plots of every pair of factors. An example for one two factor plot is given below. The plot shows the impact each pair of factors has on the response variable. This type of chart provides insights into which interactions might be significant. Interactions might be significant if the lines on the chart are not parallel.



All Factors Residual Info Worksheet

This worksheet contains a table with the residual information. The output from this example is shown below. The columns are explained below.

Residuals Data and Charts for Nitride Etch

Standard Run Order	Actual Run Order	Observed Value	Predicted Value	Residuals	Leverage	Standardized Residuals	Internally Studentized Residuals	Externally Studentized Residuals	DFFITS	Cook's Distance
1	9	550	577	-27	0.500	-0.569	-0.805	-0.785	-0.785	0.0809
1	14	604	577	27	0.500	0.569	0.805	0.785	0.785	0.0809
2	7	650	659.5	-9.5	0.500	-0.200	-0.283	-0.266	-0.266	0.0100
2	10	669	659.5	9.5	0.500	0.200	0.283	0.266	0.266	0.0100
3	2	633	617	16	0.500	0.337	0.477	0.452	0.452	0.0284
3	8	601	617	-16	0.500	-0.337	-0.477	-0.452	-0.452	0.0284
4	3	642	638.5	3.5	0.500	0.074	0.104	0.098	0.098	0.0014
4	6	635	638.5	-3.5	0.500	-0.074	-0.104	-0.098	-0.098	0.0014
5	4	1037	1044.5	-7.5	0.500	-0.158	-0.223	-0.210	-0.210	0.0062
5	15	1052	1044.5	7.5	0.500	0.158	0.223	0.210	0.210	0.0062
6	12	749	808.5	-59.5	0.500	-1.254	-1.773	-2.128	-2.128	0.3929
6	13	868	808.5	59.5	0.500	1.254	1.773	2.128	2.128	0.3929
7	1	1075	1069	6	0.500	0.126	0.179	0.168	0.168	0.0040
7	5	1063	1069	-6	0.500	-0.126	-0.179	-0.168	-0.168	0.0040
8	11	729	794.5	-65.5	0.500	-1.380	-1.952	-2.522	-2.522	0.4762
8	16	860	794.5	65.5	0.500	1.380	1.952	2.522	2.522	0.4762

Notes:
Any values that fail the following are colored in red and could be outliers.
- Leverage > 2p/n
- Standardized, internally standardized, externally standardize residuals outside the range of -3 to 3
- Absoule value DFFITS > 2Sqrt(p/n)
- Cook's Distance > 1

where p is the number of regressor variables (including b0) and n is the number of observations

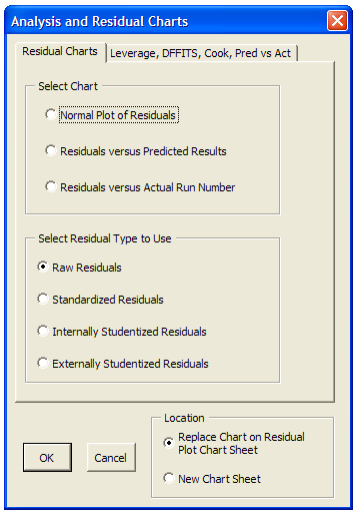
- *Standard Run Number:* the run number from the design table
- *Actual Run Number:* the order the run was actually run during the experiment
- *Observed Value:* the value of the response variable for the run
- *Predicted Value:* the value of the response variable predicted from the model
- *Residual:* the difference between the observed value and the predicted value

- *Leverage:* the amount of leverage (influence) the run has on the predicted value; the leverage values are obtained from the diagonal element of the hat matrix (see above); if the leverage for a run is greater than 2p/n, then this run is a high-leverage point and should be investigated further; p is the number of terms in the model and n is the number of runs
- *Standardized Residuals:* provides a rough check for outliers; determined by dividing each residual by the square root of the mean square error; any value outside ± 3 is a possible outlier
- *Internally Studentized Residuals:* take into account the inequality of variances across the factor space, any value outside ± 3 is a possible outlier, defined as:
  - $r_i = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$, where σ^2 is the MSE
- *Externally Studentized Residuals*: uses a different estimate of σ^2 than MSE in the above equation; estimates σ^2 based on a data set with the ith obversation removed; uses $S^2_{(i)}$, defined as:
  - $S^2_{(i)} = \frac{(n-p)MSE - e_i^2/(1-h_{ii})}{n-p-1}$

  The externally studentized residual is the defined as;
  - $t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}}$

  Any value outside ± 3 is a possible outlier.
- *DFFITS:* measures the deletion influence of run i; if absolute values is greater than $2\sqrt{p/n}$, the run is influential
  - $DFFITS_i = t_i\sqrt{h_{ii}/(1-h_{ii})}$
- *Cook's Distance:* indicates the difference between the calculated β values and the values one would have obtained, had a run been excluded; all distances should be of about equal magnitude; if not, then there is reason to believe that the run biased the estimation of the regression coefficients; values greater than 1 are influential; defined as the following:
  - $D_i = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}$

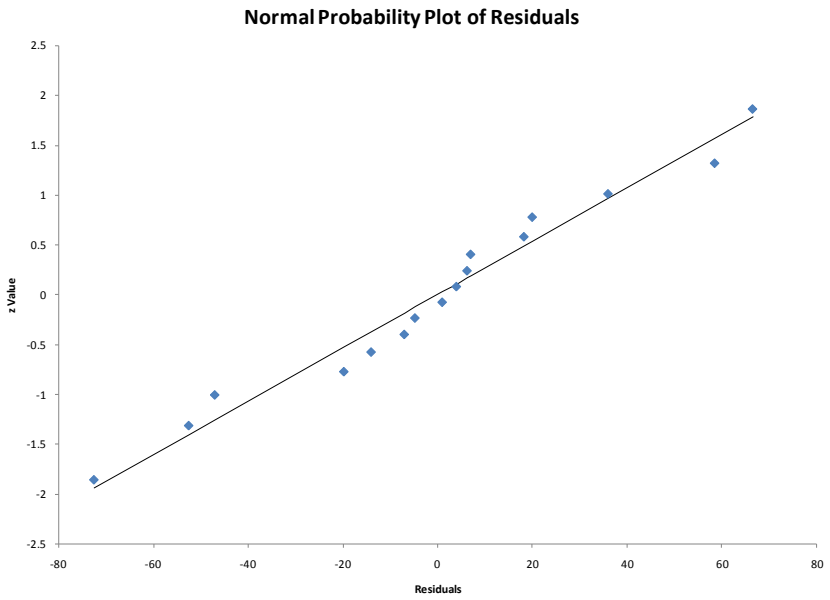
Residual Plots Sheet



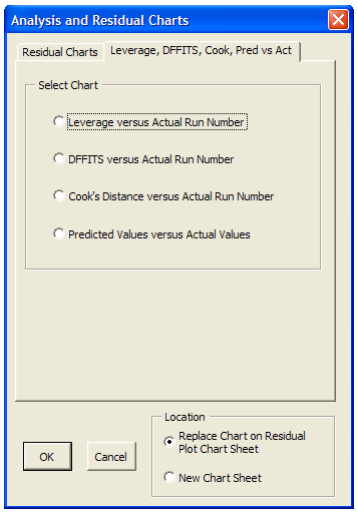
This sheet contains the residuals plot with the initial chart being the normal probability plot of residuals show on the next page. This chart is just one of many that can be generated. To access the other charts, select the "Other Charts" button that appears on the chart page. You will see the form here.

The first page of this form shows the charts that are available for the residual charts. There are three basic residual charts: normal plot of residuals, the residuals versus predicted results, and the residuals versus actual run number. Select which chart you want and then select one of the four residuals to use in the chart: raw, standardized, internally studentized, or externally studentized.

You also have the option to replace the existing chart on the Residuals Plot sheet or to have the chart placed a new sheet. If the chart is placed on a new sheet, you must come back to the Residuals Plot sheet if you want to generate additional charts.



The second page of the dialog box contains the other chart options. These include leverage, DFFITs and Cook's Distance versus the actual run number as well as the predicted value versus the actual values.

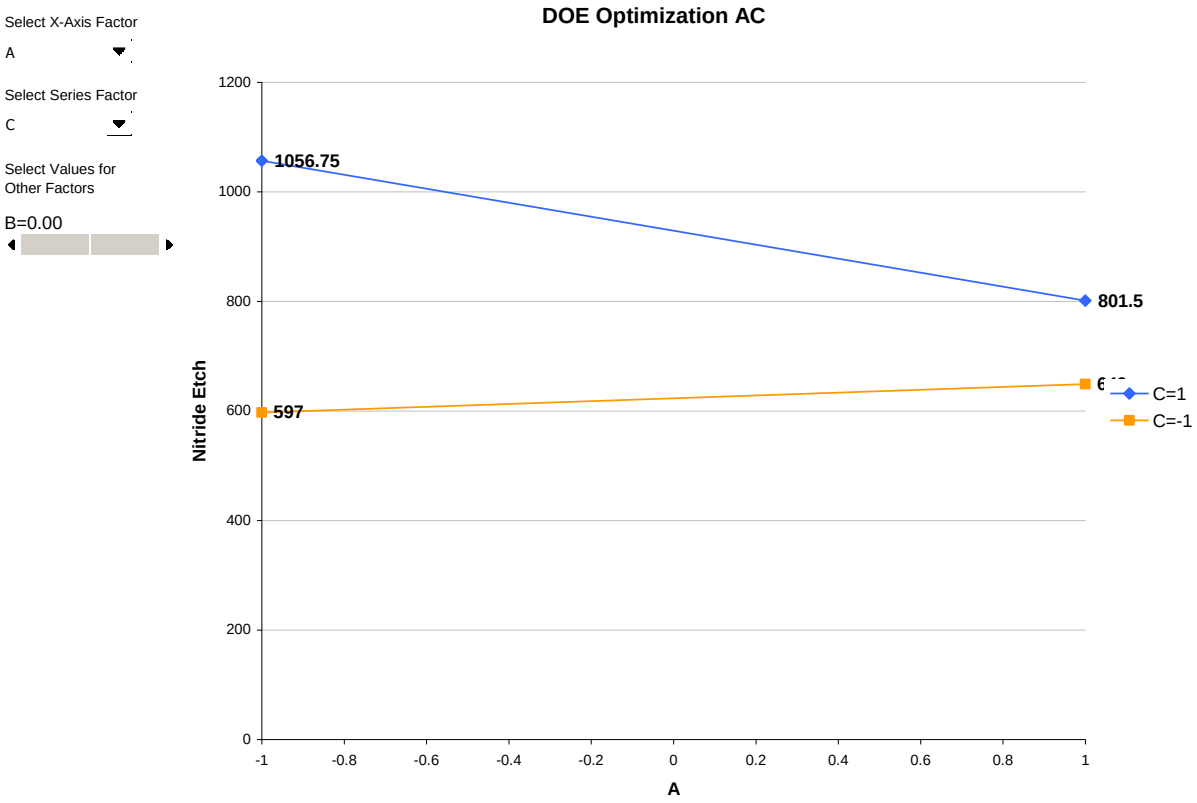


DOE Optimization

This sheet contains a chart where you can easily see the impact of changing the value of the variables on the response factor. The coded factors are used. This sheet will work for up to six factors. The optimization chart is shown below. You make changes to the chart using the boxes in the upper left-hand corner. These are described below.

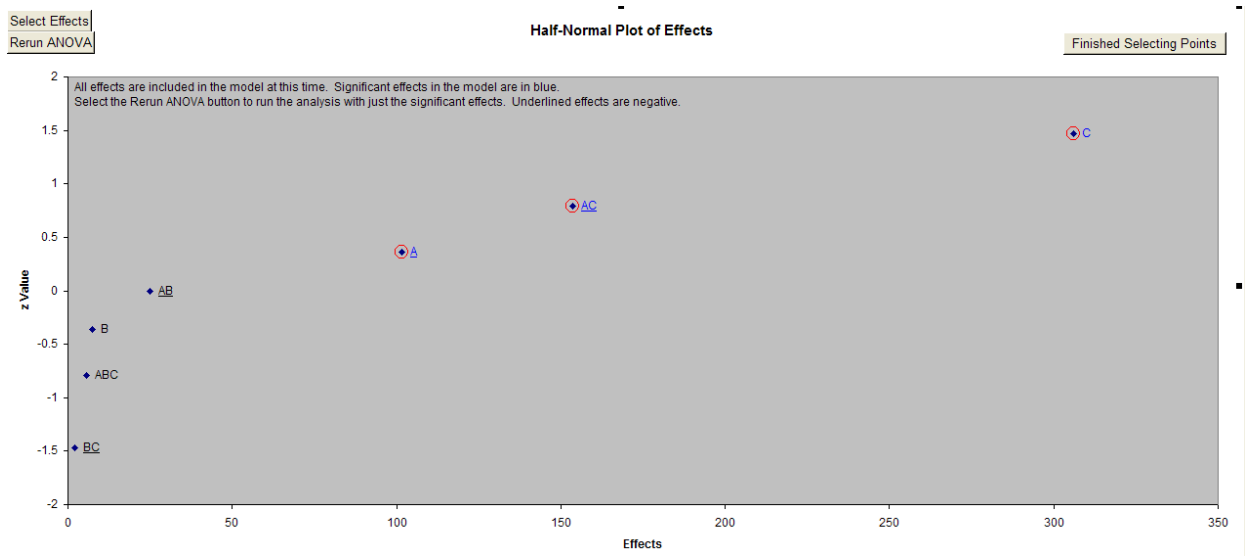
- Select X-Axis Factor: this selects the factor to be displayed on the X axis.
- Select Series Factor: this selects the factor for the lines to be drawn on the chart for the factor levels of -1 and 1.

- Select Values for Other Factors: up to four factors will be listed here; use the scroll to change the value of a factor; default value is 0.



Adding/Removing Terms in the Model for Two Level Experimental Designs

The worksheet named "Half-Normal Plot of Effects" is used to add or remove terms to the model. When the program first runs, it performs the analysis for a model that includes all the factors. To change the terms in the model, go to this worksheet and select the "Select Factors" button in the upper left hand corner of the chart. The chart will appear as shown below.



Each term that was found significant in the original analysis that included all terms are circles. You can add more terms to the terms by selecting the point on the chart. You can also remove terms from the model by selecting the circled points.

When you have finished adding or removing terms from the model, select the Finished Selecting Points in the upper right hand side of the chart. Then select the Rerun ANOVA button. This will rerun the analysis for the terms you selected.

The program will add two new worksheets called "Current Model Analysis" and "Model Residuals" and change the Residuals Plots sheet. The worksheet "Current Model Analysis" contains the same information of the "All Factors Analysis" worksheet except that is for the terms in the model. The design table is not included. The output for this worksheet in the current example is shown below. The "Model Residuals" worksheet contains the same information as the "All Factors Residual Info" worksheet except that is based on the terms in the current model. The "Residuals Plots" sheet now contains the information for the current model.

If you return to the "Half-Normal Plot of Effects" sheet, select different terms and rerun the ANOVA analysis, the three worksheets above will be changed to reflect the terms you included in the model.

Analysis Results for Nitride Etch (Y = 776.0625 + -50.8125A + 153.0625C + -76.8125AC)

ANOVA Table Based on Selected Factors and Interactions

Source	SS	df	MS	F	P-value	% Cont
A	41310.56	1	41310.56	23.76703	**0.000382**	7.77%
C	374850.1	1	374850.1	215.6609	**4.95E-09**	70.54%
AC	94402.56	1	94402.56	54.31222	**8.62E-06**	17.76%
Error	20857.75	12	1738.146			3.92%
Total	531420.9	15				100.00%

ANOVA for Model

Source	SS	df	MS	F	P-value
Model	510563.2	3	170187.7	97.91338	1.05E-08

Average	776.0625
Standard Deviation	41.69108
Coefficient of Variation	5.372129
R Squared	0.960751
Adjusted R Squared	0.950939
PRESS	37080.44
R Squared Prediction	0.930224

Factor Information

Factor	Coeff	Degrees of Freedom	Standard Error	95% Lower	95% Upper
Intercept	776.0625	1	10.42277	753.3532	798.7718
A	-50.8125	1	10.42277	-73.5218	-28.1032
C	153.0625	1	10.42277	130.3532	175.7718
AC	-76.8125	1	10.42277	-99.5218	-54.1032

Model Containing Selected Factors

Y = 776.0625 + -50.8125A + 153.0625C + -76.8125AC

Y=	776.0625	Intercept
+	-50.8125	A
+	153.0625	C
+	-76.8125	AC

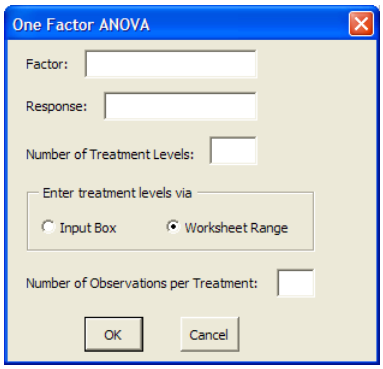
Setting Up the One and Two Factor Analysis of Variance

The one factor and two factor Analysis of Variance (ANOVA) programs are accessed through the DOE icon on the SPC menu. The setup is essentially the same for the two. The one factor setup is described here. To understand how this design is set up, we will use another example from Montgomery's book on experimental design. This example involves a plasma etching process. It is designed to determine the relationship between the RF power setting and the etch rate for single-wafer plasma etching tool. Four treatment levels of RF factor were tested (160, 180, 200, and 220). There were five observations at each treatment level.

To set up a one factor experiment, select the DOE button and select the One Factor ANOVA option. The dialog box below will appear.

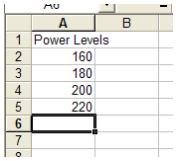
Enter the following:

- *Factor:* the name of the factor being investigated (e.g., RF Power)
- *Response:* the name of the response variable (e.g., etch rate)
- *Number of Treatment Levels:* the number of different levels for the factor (e.g., 4)
- *Enter Treatment Levels Via:* you have the option to enter the treatment levels one at a time using an Input Box or to have the levels already entered in a worksheet
- *Number of Observations per Treatment:* the number of runs within each treatment; must be equal for all treatment levels (e.g., 5)

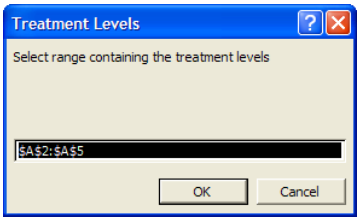


If you selected the Input Box option to enter the treatment levels, you will get the input box below for each treatment level.

It is usually easier to enter the treatment levels into a worksheet as shown below.



	A	B
1	Power Levels	
2	160	
3	180	
4	200	
5	220	
6		
7		



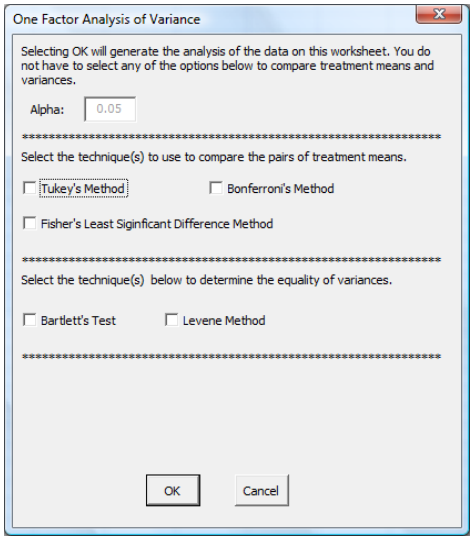
In this case, you select the worksheet range option above and you will get the dialog box above. Enter the range containing the levels and select OK. The program will generate the worksheet shown below.

One Factor ANOVA for the Effect of Power on Etch				
		Factor:	Power	
		Response:	Etch Rate	
		Treatment Levels:	4	
		Observations/Treatment Level:	5	
Actual Run Number	**Power**	**Result**		
1	220	725		
2	160	575		
3	200	600		
4	200	651		
5	180	565		
6	220	700		
7	160	542		
8	180	593		
9	200	610		
10	160	530		
11	200	637		
12	180	590		
13	220	715		
14	180	579		
15	220	685		
16	180	610		
17	160	539		
18	220	710		
19	160	570		
20	200	629		

The runs have been randomized for you. Run the experiments in the order listed and enter the result in the column marked "Result." This has already been due in the above worksheet. You are now ready to analyze the results.

Analysis of Variance Output: Calculations and Interpretation

Once the experimental design has been run and the results entered into the worksheet as shown above, you are ready to analyze the results. This process is started by selecting the DOE icon on the SPC for Excel menu. You should be on the worksheet containing the experimental results. You will see the form shown here. This form contains the options for comparing treatment means and variances. You do not have to select any of these options.



- *Alpha*: the confidence coefficient; default is 0.05.
- *Tukey's Method*: uses Tukey's method for comparing treatment means
- *Bonferroni's Method*: uses Bonferroni's method for comparing treatment means
- *Fisher's Least Significant Difference Method*: uses Fisher's LSD method for comparing treatment means
- *Bartlett's Test*: uses Bartlett's test for comparing the equality of variances
- *Levene's Method*: uses Levene's Method for comparing the equality of variances.

More information on each of these techniques is given in the General Instruction Manual for SPC for MS Excel Version 4.0. If you select on the methods for comparing treatment means, an additional chart will be generated showing the treatments. With two factors, if you apply these tests at the ab level or at the abc level. Additional mean charts will be generated for each level.

When you select OK, the data is moved to a new workbook and the results generated. The following worksheets are in the new workbook:

- Raw Data: this worksheet contains the raw data from the previous workbook
- ANVOA: this worksheet contains the results for the following:
  - Table of results by treatment
  - ANOVA table for the factor(s)
  - ANOVA table for the model
  - Average, standard deviation, coefficient of variation, R^2, adjusted R^2, PRESS and R^2 prediction
  - Table of treatment means, standard errors and 95% upper and lower confidence intervals
  - Any select comparison of treatment means and/or variances
- Residuals Plots: contains the normal probability plot of the raw residuals initially but has options for many more
- All Factors Residuals Info: contains the following:
  - Standard run number
  - Actual run number
  - Observed value
  - Predicted value
  - Raw residuals
  - Leverage
  - Standardized residuals
  - Internally studentized residuals
  - Externally studentized residuals
  - DFFITS
  - Cook's distance

Details on each worksheet (except for Raw Data worksheet which contains the raw data) are given below for the one factor ANOVA. The two factor ANOVA output is very similar.

ANOVA Worksheet

The ANOVA worksheet the basic results from the analysis. Each part of the output is described below.

Table of Results

The top part of the worksheet displays the data with the levels of the factor in column A and the observations in the adjacent columns.

	Observations				
Power	**1**	**2**	**3**	**4**	**5**
160	575	542	530	539	570
180	565	593	590	579	610
200	600	651	610	637	629
220	725	700	715	685	710

ANOVA Table

The ANOVA table for the factor is given next as shown below. The table contains the source of variation, the sum of squares, degrees of freedom, mean square, F value and p value. If the p value is ≤ 0.05, it is in red.

ANOVA for Power					
Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p Value
Power	66870.550	3	22290.183	66.797	0.0000
Error	5339.200	16	333.700		
Total	72209.750	19			

The columns of the ANOVA table are:

- Source: the source of variation, which includes the factor (Power in this example), the error, and the total
- Sum of Squares: the sum of squares for each source of variation; the treatment sum of squares below corresponds to the sum of squares for the power
  - $SS_{Treatment} = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$
  - $SS_{Total} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N}$
  - $SS_{Error} = SS_{Total} - SS_{Treatment}$

  where a is the number of treatment levels, n_i is the number of observations for the i^{th} treatment level (must be constant), N is the total number of observations, and $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$ and $y_{ii} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}$.
- Degrees of Freedom: the degrees of freedom for each source of variation
  - $df_{Treatment}$ = a – 1
  - df_{Total} = n*(a-1)
  - $df_{Error} = df_{Total} - df_{Treatment}$
- Mean Square: the mean square for the source is the variance associated with that source and is determined by dividing the source sum of squares by the degrees of freedom for that source.
  - MS = SS/df
- F: value from the F distribution
  - The F Value for a source of variation is used to compare the variance associated with that source with the error variance.
    - F = MS/MSE where MS is the mean square for a source and MSE the mean square error
- p-Value: the probability value that is associated with the F Value for a source of variation
  - It represents the probability of getting a given F Value if the source does not have an effect on the response.
  - If the p-value is <=0.05, it is considered to have a significant effect on the response.
  - A p-value above 0.20 is not considered to have an effect on the response
  - If the p-value is between 0.05 and 0.20, it or may not have a significant effect.

ANOVA Table for the Model

The next part of the worksheet contains the ANOVA table for the model. The output from this example is shown below. The columns in the ANOVA table have been explained above; the other parameters are explained in the DOE section of this manual.

ANOVA for Model					
Source	SS	df	MS	F	p Value
Model	66870.550	3	22290.183	66.797	0.0000
Average		617.750			
Standard Deviation		18.267			
Coefficient of Variation		2.957			
R Squared		92.61%			
Adjusted R Squared		91.22%			
PRESS		8342.5			
R Squared Prediction		88.45%			

Treatment Mean Confidence Intervals

The next table provides the 95% confidence intervals for the treatment means. The output for this example is shown below. The standard error the square root of the MSE/n where n is the number of observations per treatment.

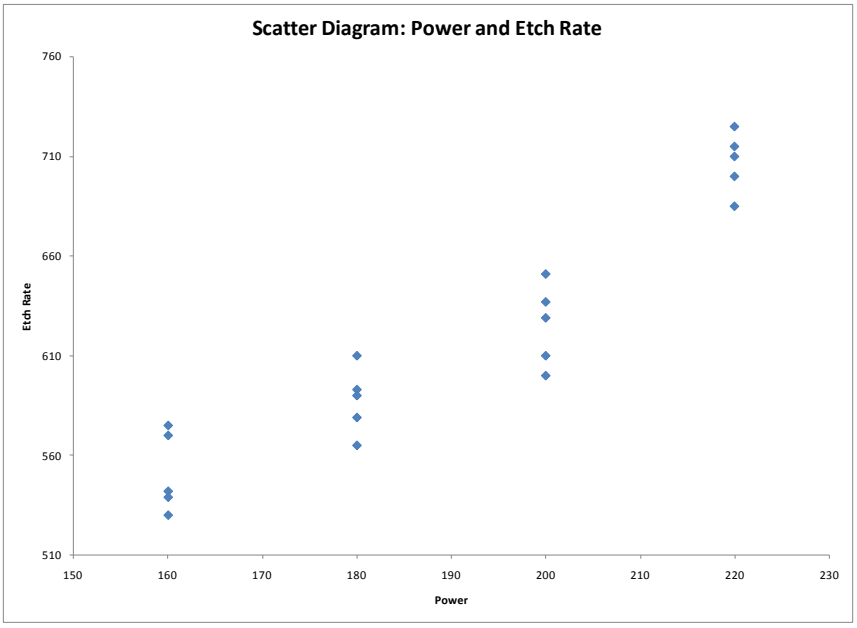
Treatment	Mean	Stand. Err.	95% LCon	95% UCon
160	551.200	8.169	533.882	568.518
180	587.400	8.169	570.082	604.718
200	625.400	8.169	608.082	642.718
220	707.000	8.169	689.682	724.318

Residuals and Residual Plots Sheets

These two worksheets contain the residual analysis and the associated plots. This output is explained in the DOE section above (click here).

Scatter Sheet

This chart is a scatter diagram of the results. The scatter diagram for this data is shown below.



Setting Up the Multiple Linear Regression

We will use an example from Montgomery's regression book.[1] An engineer employed by a soft drink beverage bottler is analyzing what impacts delivery times. He decides that two factors that impact the time could be the number of cases a driver delivers as well as how far the driver has to walk at the customer's facility. He has collected 25 observations for delivery time (minutes), the number of cases, and distance walked (feet). The data is shown below. We want to use this data to determine if either factor impacts delivery time and if we can build a model to predict delivery time. The steps below show how to do this using the SPC for MS Excel software. In this example, we are using the following model:

$$Y = b_o + b_1x_1 + b_2x_2$$

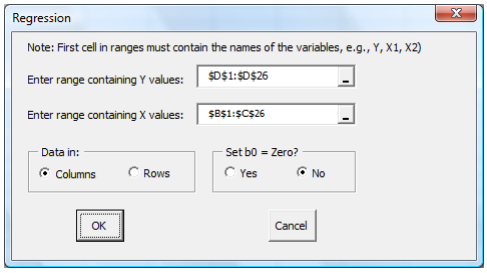
where Y = response variable (delivery time)
x_1 = predictor 1 (number of cases)
x_2 = predictor 2 (distance walked)
b_o = intercept
b_1 = coefficient for predictor 1
b_2 = coefficient for predictor 2

1. Enter the data into a spreadsheet as shown below.

Observation	Number of Cases	Distance	Delivery Time
1	7	560	16.68
2	3	220	11.5
3	3	340	12.03
4	4	80	14.88
5	6	150	13.75
6	7	330	18.11
7	2	110	8
8	7	210	17.83
9	30	1460	79.24
10	5	605	21.5
11	16	688	40.33
12	10	215	21
13	4	255	13.5
14	6	462	19.75
15	9	448	24
16	10	776	29
17	6	200	15.35
18	7	132	19
19	3	36	9.5
20	17	770	35.1
21	10	140	17.9
22	26	810	52.32
23	9	450	18.75
24	8	635	19.83
25	4	150	10.75

1. [1] Montgomery, D. C., Peck, E., and Vining, G. G, Introduction to Linear Regression Analysis, 4th Edition, John Wiley & Sons, 2006

2. Select the shaded area as shown above (the data and the headings; not the observation numbers)
3. Select the multiple linear regression icon ( ) from the SPC menu.
4. Fill in the Regression form.
    a. *Enter range containing the Y values:* enter the worksheet range containing the Y values; default is the last column (or row) selected on the worksheet; include the name of the response variable.
    b. *Enter range containing the X values:* enter the worksheet range containing the X (predictor) values; the default is the range selected minus the last column (or row) on the worksheet; include the name of the predictor values.
    c. *Data in:* the data can be in columns or rows; the default is based on the number of columns and rows selected on the worksheet.
    d. *Set b0 = Zero:* option to set the intercept to zero; default it no
5. Select OK.
6. The regression analysis is performed. A new workbook is created and the output placed in the new workbook. The output from the regression analysis is shown below.



Multiple Linear Regression Output

The multiple linear regression analysis provides the following worksheets with the associated output:

- Data: the original data
- Original Regression Summary
    - ANOVA table for the model
    - Coefficients table
        - Coefficient
        - Standard Error
        - t statistic
        - 95% upper and lower confidence interval for the coefficient
        - VIF (variation inflation factor)
        - Standardized coefficients
    - Regression statistics
        - R
        - R squared
        - Adjusted R squared
        - Mean
        - Standard error
        - Coefficient of variation
        - Observations
        - Durbin-Watson statistic
        - PRESS
        - R squared prediction
- Original Regression Data
    - Observation Number
    - Observed Value
    - Predicted Value
    - Raw residuals
    - Leverage

- Standardized residuals
- Internally studentized residuals
- Externally studentized residuals
- DFFITS
- Cook's distance
- Standard error of estimated mean
- 95% lower and upper confidence limits

- Residual Plots
  - Residual plots for each type of residual (raw, standardized, internally studentized, externally studentized)
    - Normal plot of residuals
    - Residuals versus predicted results
    - Residuals versus actual run number
    - Residuals versus predictor (X) variable
  - Other plots
    - Leverage versus actual run number
    - DFFITs versus actual run number
    - Cook's distance versus actual run number
    - Predicted values versus predicted values

Each worksheet in the output (except the Data worksheet which contains the raw data) is discussed below. The options for updating the regression analysis are then given.

Original Regression Summary Worksheet

This worksheet contains a summary of the regression analysis. The output is shown below.

Regression Summary for Delivery Time

Revise

ANOVA Table

	df	SS	MS	F	p value
Model	2	5550.811	2775.405	261.24	0.0000
Residual	22	233.732	10.624		
Total	24	5784.543			

Coefficients

	Coefficient	Standard Error	t Stat	p Value	95% Lower	95% Upper	VIF	Stand. Coeff
Intercept	2.341	1.097	2.135	0.0442	0.067	4.616		
Number of Cases	1.616	0.171	9.464	0.0000	1.262	1.970	3.12	0.716
Distance	0.0144	0.00361	3.981	0.0006	0.0069	0.0219	3.12	0.301

Regression Statistics

R	97.96%
R Square	95.96%
Adjusted R Square	95.59%
Mean	22.384
Standard Error	3.259
Coefficient of Variance	14.562
Observations	25
Durbin-Watson Statistic	1.170
PRESS	459.039
R Squared Prediction	92.06%

ANOVA Table

The top part of this worksheet contains the ANOVA Table for the regression. This output has been explained in the ANOVA section above. The p value will be printed in red if it is less than 0.05. This means that the regression is statistically significant.

Coefficients

This table contains the results for the coefficients and the intercept (if that option was selected). The table contains the following:

- *Coefficient:* these are intercept (if that option was selected) and the coefficients for the predictor factors; the coefficients are found using matrix algebra:
  - $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
- *Standard Error:* the standard error of the coefficient defined by:
  - se(β)=$\sqrt{\sigma^2 C_{ii}}$ where σ^2 is the mean square residual or error and C_{ii} is the diagonal elements of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix.
- *t Stat:* the t statistic which is the coefficient divided by the standard error
- *p Value:* the p value associated with getting the t statistic
- 95% Lower and Upper: the lower and upper confidence intervals defined as:
  - β±t(se(β)) where t is the value of the t distribution for 0.05 and the residual degrees of freedom,
- VIF: the variance inflation factors which measure the multi-collinearity (the correlation between predictor); VIF is the diagonal elements of the $(\boldsymbol{W}'\boldsymbol{W})^{-1}$ matrix (which uses unit length scaling);
  - VIF = 1, no correlation
  - 1 < VIF < 5, moderate correlation
  - 5 < VIF < 10, high correlation
  - VIF > 10, may be impacting the regression analysis
- Stand. Coeff: the standardized regression coefficients; these are dimensionless coefficients that give you an estimate of the relative impact of each coefficient on the response variable;
  - $\boldsymbol{b} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{y}^{\boldsymbol{0}}$

Please see Montgomery's book on regression for more information on these calculations.

Regression Statistics

The following regression statistics are given:

- *R Squared:* measures the proportion of the total variability measured explained by the model
  - $R^2 = 1 - \frac{SS_{Residual}}{SS_{Model} + SS_{Residual}}$
- *Adjusted R Squared:* the value of R^2 adjusted for the size of the model (the number of factors in the model)
  - $R^2_{Adj} = 1 - \frac{SS_{Residual}/df_{Residual}}{\left(\frac{SS_{Model}+SS_{Residual}}{df_{Model}+df_{Residual}}\right)}$
- *Mean:* the average of all the responses
- *Standard Error*: the square root of the mean square residuals
- *Coefficient of variation:* the error expressed as a % of the mean, 100(Standard Error/Mean)
- *Observations:* the number of observations
- *Durbin-Watson Statistic:* this is a measure of the autocorrelation in the residuals 9; if the residuals are correlated, the predictor factors may appear significant when they are not because the standard error of the coefficients is underestimated; the equation for the statistic is given below where e = residual
  - $d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$
- *PRESS:* Predicted Error Sum of Squares is a measure of how well the model will predict new values and is given below where e_i is the ith residual and h_{ii} is the diagonal element of the hat matrix (H=X(X'X)$^{-1}$X')

- $PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2$
- *R Squared Prediction*: indication of the predictive capability of the model; the percent of the variability the model would be expected to explain with new data
  - $R^2 = 1 - \frac{PRESS}{SS_{Total}}$

Original Residuals Data Worksheet

This worksheet contains a table with the original residual information. The output from this example is shown below. The columns are explained below.

Observation Number	Observed Value	Predicted Value	Residuals	Leverage	Standardized Residuals	Internally Studentized Residuals	Externally Studentized Residuals	DFFITS	Cook's Distance	Standard Error of Estimated Mean	95% Lower CL	95% Upper CL
1	16.68	21.708	-5.028	0.102	-1.543	-1.628	-1.696	-0.571	0.100	1.040	19.551	23.865
2	11.5	10.354	1.146	0.071	0.352	0.365	0.358	0.099	0.003	0.867	8.556	12.151
3	12.03	12.080	-0.050	0.099	-0.015	-0.016	-0.016	-0.005	0.000	1.024	9.956	14.204
4	14.88	9.956	4.924	0.085	1.511	1.580	1.639	0.501	0.078	0.952	7.981	11.931
5	13.75	14.194	-0.444	0.075	-0.136	-0.142	-0.139	-0.039	0.001	0.893	12.343	16.046
6	18.11	18.400	-0.290	0.043	-0.089	-0.091	-0.089	-0.019	0.000	0.675	17.000	19.799
7	8	7.155	0.845	0.082	0.259	0.270	0.265	0.079	0.002	0.932	5.222	9.089
8	17.83	16.673	1.157	0.064	0.355	0.367	0.359	0.094	0.003	0.823	14.967	18.380
9	79.24	71.820	7.420	0.498	2.276	3.214	4.311	4.296	3.419	2.301	67.049	76.592
10	21.5	19.124	2.376	0.196	0.729	0.813	0.807	0.399	0.054	1.444	16.129	22.119
11	40.33	38.093	2.237	0.086	0.686	0.718	0.710	0.218	0.016	0.957	36.109	40.076
12	21	21.593	-0.593	0.114	-0.182	-0.193	-0.189	-0.068	0.002	1.099	19.314	23.872
13	13.5	12.473	1.027	0.061	0.315	0.325	0.318	0.081	0.002	0.806	10.802	14.144
14	19.75	18.682	1.068	0.078	0.328	0.341	0.334	0.097	0.003	0.912	16.792	20.573
15	24	23.329	0.671	0.041	0.206	0.210	0.206	0.043	0.001	0.661	21.958	24.699
16	29	29.663	-0.663	0.166	-0.203	-0.223	-0.218	-0.097	0.003	1.328	26.909	32.417
17	15.35	14.914	0.436	0.059	0.134	0.138	0.135	0.034	0.000	0.795	13.266	16.562
18	19	15.551	3.449	0.096	1.058	1.113	1.119	0.365	0.044	1.011	13.454	17.649
19	9.5	7.707	1.793	0.096	0.550	0.579	0.570	0.186	0.012	1.012	5.607	9.806
20	35.1	40.888	-5.788	0.102	-1.776	-1.874	-1.997	-0.672	0.132	1.039	38.732	43.044
21	17.9	20.514	-2.614	0.165	-0.802	-0.878	-0.873	-0.389	0.051	1.325	17.766	23.262
22	52.32	56.007	-3.687	0.392	-1.131	-1.450	-1.490	-1.195	0.451	2.040	51.777	60.236
23	18.75	23.358	-4.608	0.041	-1.414	-1.444	-1.482	-0.308	0.030	0.662	21.984	24.731
24	19.83	24.403	-4.573	0.121	-1.403	-1.496	-1.542	-0.571	0.102	1.132	22.055	26.750
25	10.75	10.963	-0.213	0.067	-0.065	-0.068	-0.066	-0.018	0.000	0.841	9.218	12.708

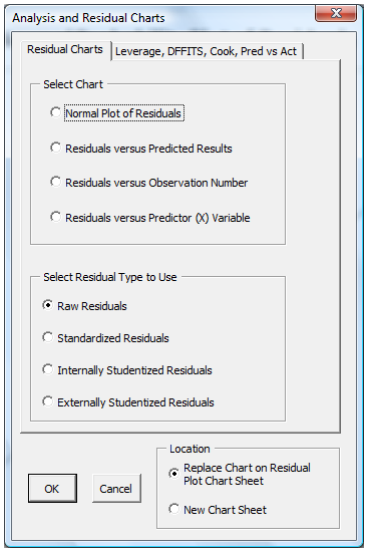
- *Observation Number:* the observation number assigned by the program
- *Observed Value:* the value of the response variable for the observation
- *Predicted Value:* the value of the response variable predicted from the model
- *Residual:* the difference between the observed value and the predicted value
- *Leverage:* the amount of leverage (influence) the run has on the predicted value; the leverage values are obtained from the diagonal element of the hat matrix (see the DOE section); if the leverage for a run is greater than 2p/n, then this run is a high-leverage point and should be investigated further; p is the number of terms in the model and n is the number of runs
- *Standardized Residuals:* provides a rough check for outliers; determined by dividing each residual by the square root of the mean square error; any value outside ± 3 is a possible outlier
- *Internally Studentized Residuals:* take into account the inequality of variances across the factor space, any value outside ± 3 is a possible outlier, defined as:
  - $r_i = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$, where σ^2 is the mean square error (MSE)
- *Externally Studentized Residuals*: uses a different estimate of σ^2 than MSE in the above equation; estimates σ^2 based on a data set with the ith obversation removed; uses $S^2_{(i)}$, defined as:
  - $S^2_{(i)} = \frac{(n-p)MSE - e_i^2/(1-h_{ii})}{n-p-1}$

  The externally studentized residual is the defined as;
  - $t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}}$

  Any value outside ± 3 is a possible outlier.

- *DFFITS:* measures the deletion influence of run i; if absolute values is greater than $2\sqrt{p/n}$, the run is influential
    - $DFFITS_i = t_i\sqrt{h_{ii}/(1-h_{ii})}$
- *Cook's Distance:* indicates the difference between the calculated β values and the values one would have obtained, had a run been excluded; all distances should be of about equal magnitude; if not, then there is reason to believe that the run biased the estimation of the regression coefficients; values greater than 1 are influential; defined as the following:
    - $D_i = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}$
- *Standard Error of Estimated Mean:* the stand
- *95% Lower CL:* the lower confidence limit for the predicted value
- *95% Upper CL:* the upper confidence limit for the predicted value

Residual Plots Sheet



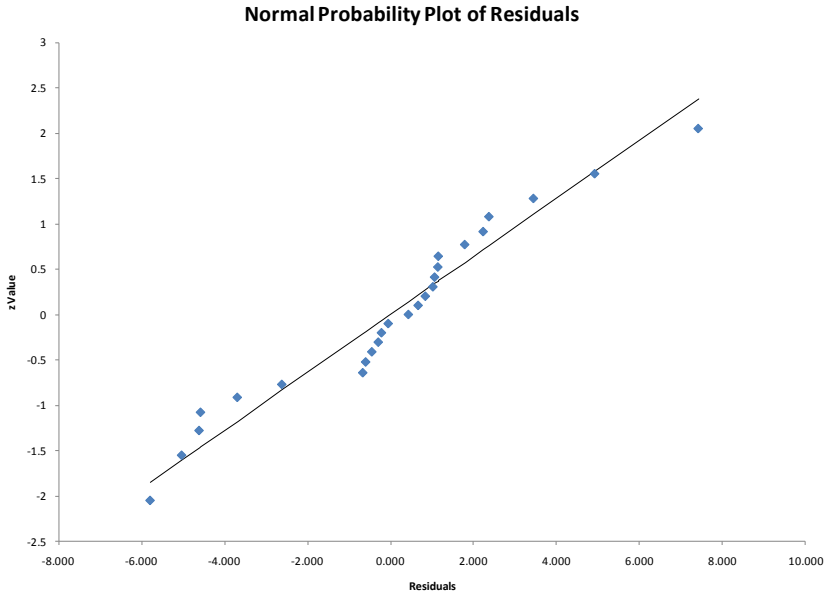
This sheet contains the residuals plot with the initial chart being the normal probability plot of residuals show on the next page. This chart is just one of many that can be generated. To access the other charts, select the "Other Charts" button that appears on the chart page. You will see the form here.

The first page of this form shows the charts that are available for the residual charts. There are four basic residual charts:

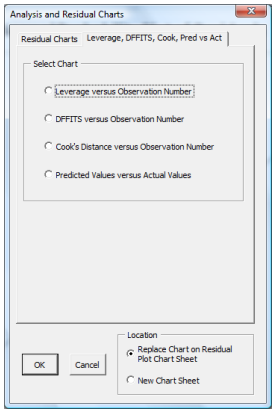
- Normal plot of residuals
- Residuals versus predicted results
- Residuals versus observation number
- Residuals versus predictor (x) variable

Select which chart you want and then select one of the four residuals to use in the chart: raw, standardized, internally studentized, or externally studentized.

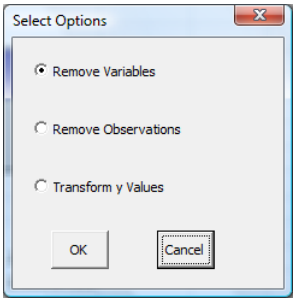
You also have the option to replace the existing chart on the Residuals Plot sheet or to have the chart placed a new sheet. If the chart is placed on a new sheet, you must come back to the Residuals Plot sheet if you want to generate additional charts.



The second page of the dialog box contains the other chart options. These include leverage, DFFITs and Cook's Distance versus the observation number as well as the predicted value versus the actual values.



Revising the Regression



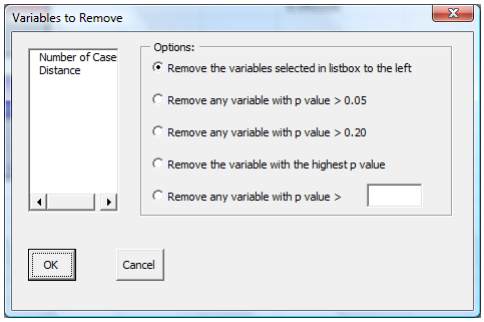
The program allows you to make changes to the regression. The changes are made from the worksheet called "Original Regression Summary." There is a button on that worksheet labeled "Revise." Select this button and you will get the form shown here. There are three options:

1. Remove Variables
2. Remove Observations
3. Transform y Values

These options are discussed below. The revisions to the regression may be in the same workbook or in a different workbook. If you are removing variables, new worksheets will be added to the existing workbook. These worksheets will be labeled starting with Current. *However, the Residuals Plot sheet will always be for the current model.* If you remove observations or transform y values, you are changing the data set and a new workbook will be created and the results stored in that workbook.

Remove Variables

If you select the remove variable option, you will see the form below. The options are:



- *Remove the variables selected in list box to the left:* select the variables you want to remove in the list box shown; you may select more than one.
- *Remove any variable with p value > 0.05*: this option removes all variables with a p value greater than 0.05.
- *Remove any variable with p value > 0.20*: this option removes all variables with a p value greater than 0.20.
- *Remove the variable with the highest p value*: this option removes the variable with the largest p value.
- *Remove any variable with p value >* _: this option removes all variables with a p value greater than a value you enter.

Select OK and new worksheets containing the current model results will be inserted into the existing worksheet.

Remove Observations

The residuals data worksheet contains information on the residuals. Some of the cells on this page may be in red representing possible outliers. The notes on the bottom of the page explain the outliers. If you want to remove some of these observations, select the Remove Observations option. You will get a box that lists all the observations. Select the observation(s) you want to delete and then select OK. A new workbook will be created with the new analysis.

Transform Y Variable

You also have the option to transform the Y (response) value using some built in transformations. If you want to transform the Y values, select that option and the form below will be shown. Select the transformation option you want and a new workbook will be created with the results.

The transformation options are:

- Square root
- Arcsine
- Log
- Reciprocal Square Root
- Reciprocal
- Box-Cox Transformation