Tab 8:

Multiple Regression

Tab 8: Multiple Regression

PURPOSE:

To introduce the multiple regression equation as a possible model for a process with more than one independent variable.

OBJECTIVES:

- Understand the components of the multiple regression equation - the constant and the coefficients ('parameters')
- Use the concept of centering to ensure that the regression model is orthogonal
- Use residual plots for evaluation of the 'goodness' of the model
- Evaluate the regression model by looking at p-values, R^2, and the Standard Deviation of the Residuals.
- Generate contour plots from the data and determine optimal conditions for the "X"s

What is Multiple Regression?

- A means of defining a relationship between a continuous "Y" variable and multiple, continuous "X" variables
- A mathematical model of the process, based upon data you provide

Why use Multiple Regression?

- It provides the ability to model the process with either a linear equation or a quadratic equation (an equation with squared terms)

What is the general form of the equation?

$$Y_i = a + b_i * X_{1i} + \dots + b_k * X_{ki} + \text{error}$$

Caution!! **As with any form of model-building, be careful about the conclusions you draw from the model. This is especially true if you are running Regression on Baseline data.**

If Regression is used with Baseline data, you MUST run a DOE to confirm the model (demonstrate that these "X"s really DO control the "Y")

‘Centering’ the “X”s to Provide Orthogonality

The “X”s in a multiple regression equation can be individual, distinct variables, or they can be related - such as X_1^2 or $X_1 * X_2$.

What if the squared term (X_1^2) is significant, but the linear term (X_1) is not? How can the effects be separated when the two terms are related? (X_1 is <u>definitely</u> related to X_1^2!)

We have to ‘transform’ the related “X” variables in order to separate their individual effects on “Y”. The transformation method is called ***‘centering’***.

<u>How do I ‘center’ data</u>?

For each “X” variable, subtract the average of that column and then compute the square: $(x_i - \bar{x})^2$

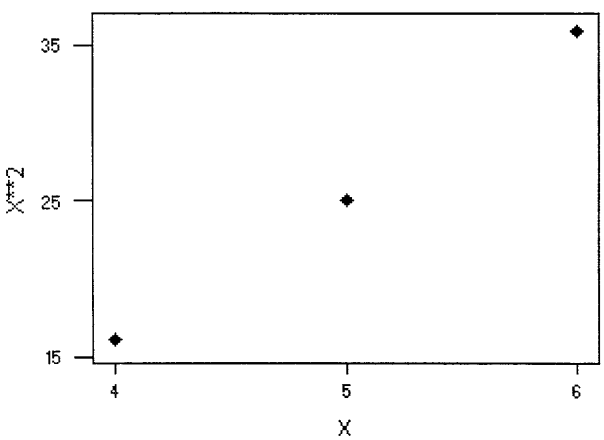
The original and centered data are nearly orthogonal - this allows effects to be separated

Let’s try an example . . . .

Centering Example:

Data set:

X	X^2
4	16
5	25
6	36



- X and X^2 change together
- The graph shows high 'correlation' (r) - a value that indicates how well the data points fall on a straight line. For this data, ***r* = .998** (with perfect correlation, r = 1.0).
- We want <u>independent</u> "X" variables: ***r* = 0**

If both X and X^2 affect the process, they can be studied **independently** by 'centering' them.

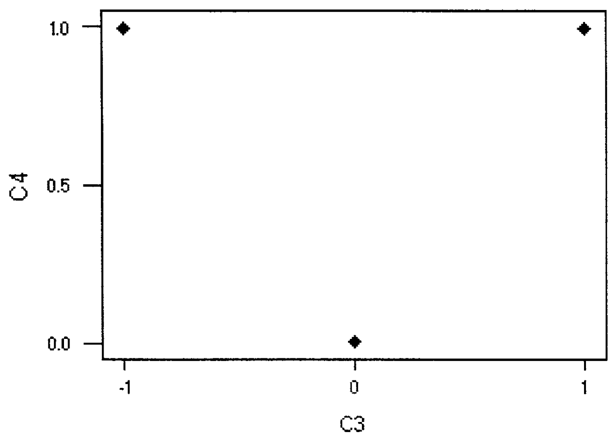
Subtract the average X value from each data point:

Average = (4 + 5 + 6) / 3 = **5**

Subtract the average value, 5, from each X data point to create 'centered' data:

(X - 5)	$(X - 5)^2$
-1	1
0	0
1	1

Graph the 'centered' data:



The correlation coefficient of the transformed data is:

r = 0.0 - the two variables are now <u>independent</u>!

(**Stat>Basic Statistics>Correlation. . .** can be used to calculate the 'r' value)

- The <u>linear</u> effect can be estimated by comparing the "Y" response at the low level of "X" to the response at the high level of "X".
- The <u>quadratic</u>, or curvature, effect can be estimated by comparing the response at the middle value of "X" to the average of the response when "X" is high and low.

Multiple Regression Example in Minitab

The following example is from page 358 of:
Richard A. Johnson. *Miller and Freund's Probability and Statistics for Engineers: Fifth Edition*. Prentice Hall. 1994.

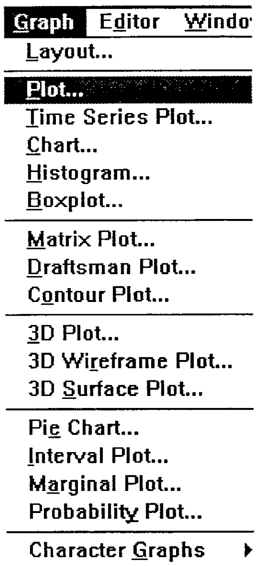
The **objective** is to estimate an equation describing the effects of :
Percent of element A, and
Percent of element B
on the number of twists required to break a forged alloy bar.

"*Twists*" is the Y (response) variable.
"***A***" and "***B***" are the X (independent) variables.

Enter "*twists*" in C1, "***A***" in C2, and "***B***" in C3.

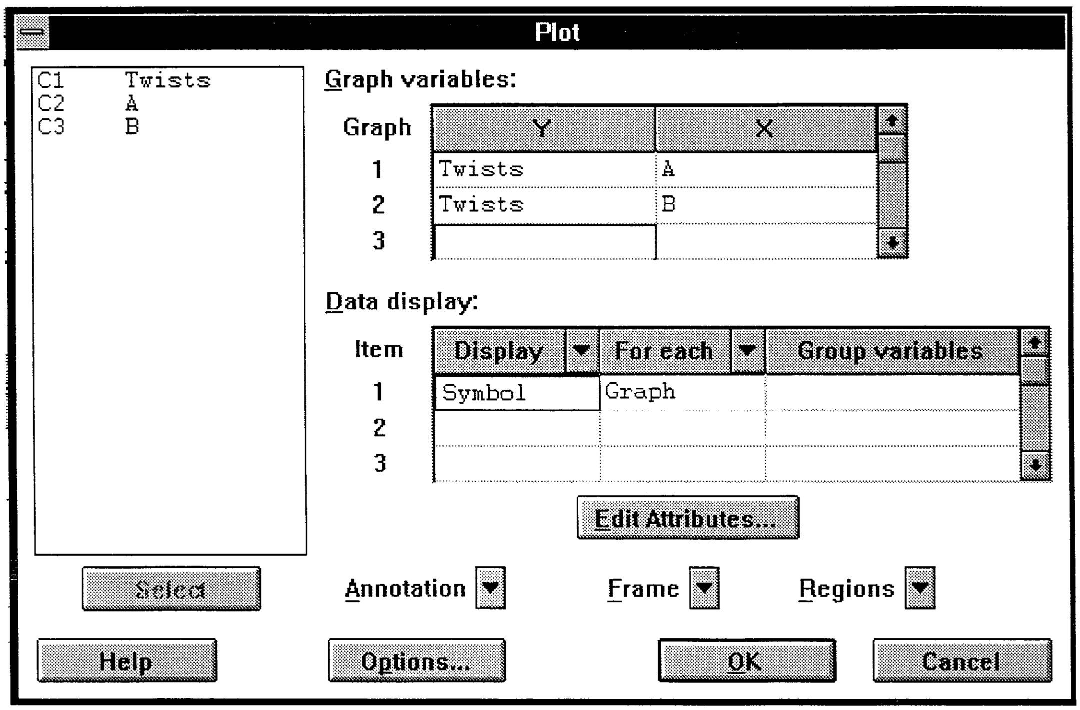
C1	C2	C3
Twists	***A***	***B***
41	1	5
49	2	5
69	3	5
65	4	5
40	1	10
50	2	10
58	3	10
57	4	10
31	1	15
36	2	15
44	3	15
57	4	15
19	1	20
31	2	20
33	3	20
43	4	20

Check for relationships between the "X"s and the "Y"



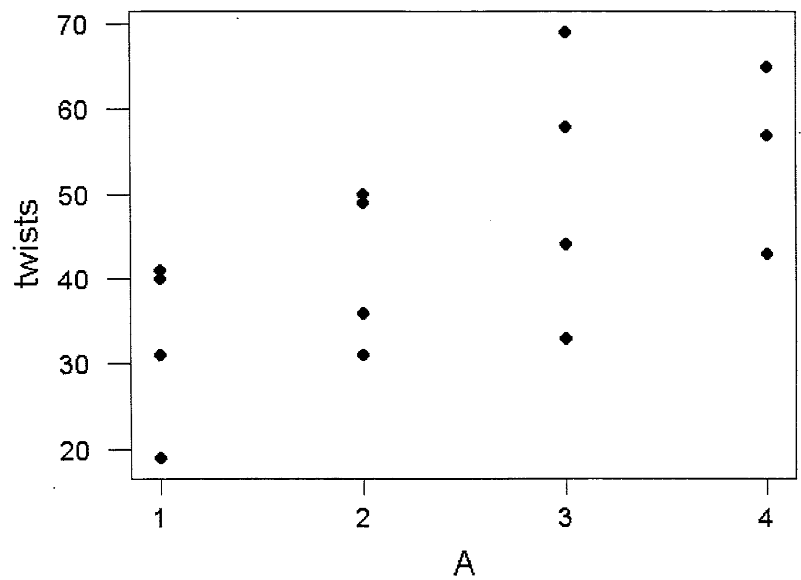
Graph **Y vs. A** and **Y vs. B** using scatterplots:
Graph>Plot

Fill in the dialog box as shown:



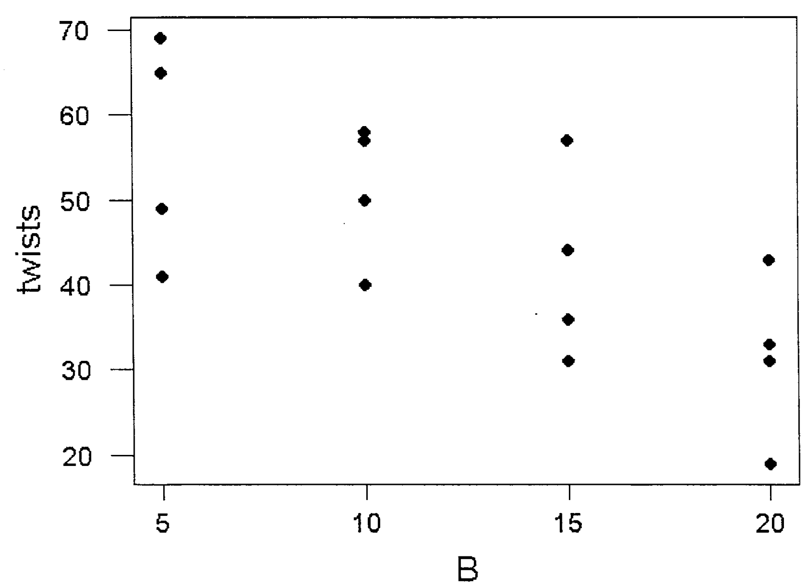
Click 'OK'

The Initial Plots



Interpretation:

- Twists increase as "A" increases
- The relationship between "Y" and "A" looks like it might be curved or linear



Interpretation:

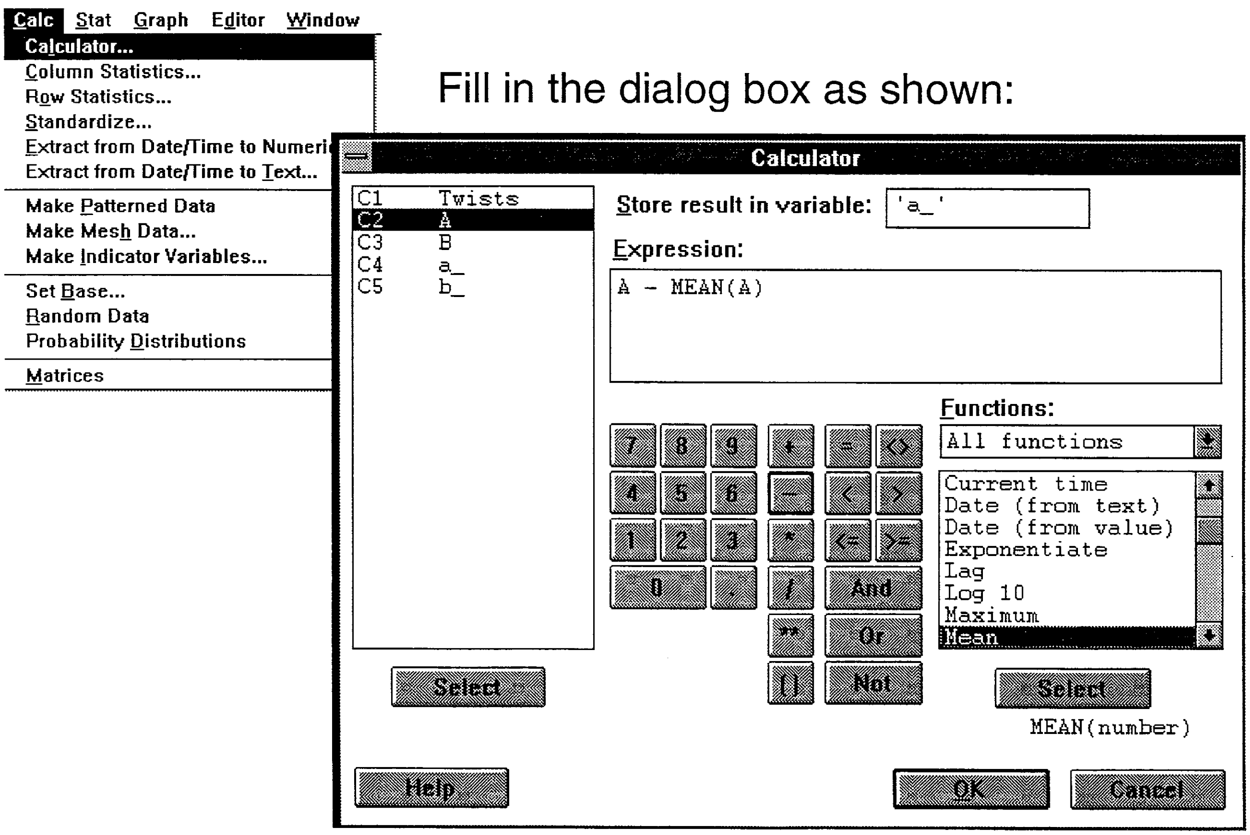
- Twists decrease as "B" increases
- The relationship looks like it might be curved or linear

Since there is a possibility that the model is non-linear, we need to center the "X" data to see the effects of any non-linear terms.

First, name the two new columns in the Data window that will contain 'centered A' data and 'centered B' data: label column C4 "a_", and label C5 "b_".

Use the Minitab calculator to create the 'centered' columns.

Fill in the dialog box as shown:



Click 'OK'

Use 'Ctrl-e' to return to this dialog box, and repeat the operation for variable "B". Click 'OK'.

Revised Minitab data window, including a_ and b_:

	C1	C2	C3	C4	C5	C6
↓	twists	A	B	a_	b_	
1	41	1	5	-1.5	-7.5	
2	49	2	5	-0.5	-7.5	
3	69	3	5	0.5	-7.5	
4	65	4	5	1.5	-7.5	
5	40	1	10	-1.5	-2.5	
6	50	2	10	-0.5	-2.5	
7	58	3	10	0.5	-2.5	
8	57	4	10	1.5	-2.5	
9	31	1	15	-1.5	2.5	
10	36	2	15	-0.5	2.5	
11	44	3	15	0.5	2.5	
12	57	4	15	1.5	2.5	
13	19	1	20	-1.5	7.5	
14	31	2	20	-0.5	7.5	
15	33	3	20	0.5	7.5	
16	43	4	20	1.5	7.5	

Performing the Multiple Regression Analysis

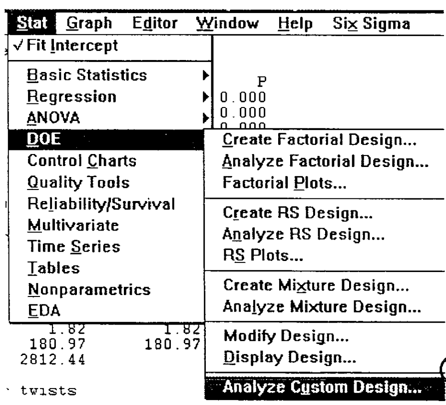
The graphs indicate a quadratic model (one with squared terms) may be most appropriate, since the graphs seem to have some curvature. We can always start with a quadratic model and simplify it later if a simpler model would give a better fit.

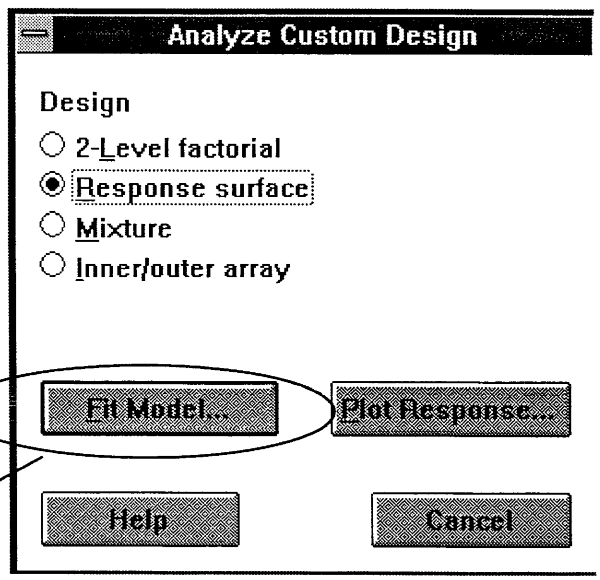
But, there is no selection under **Stat>Regression>Regression** to generate a quadratic model for more than one 'X' . . .

How do I perform Multiple Regression with squared terms in Minitab?

Use Response Surface models under DOE!

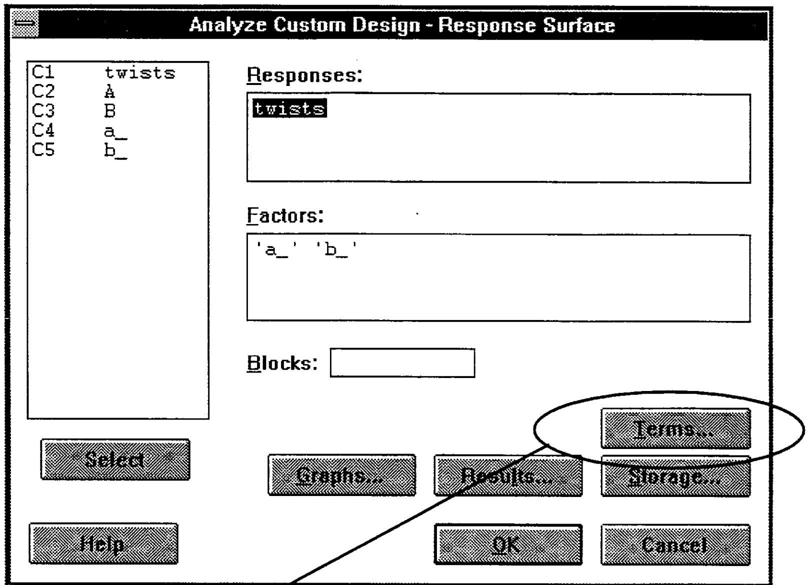
Stat>DOE>Analyze Custom Design



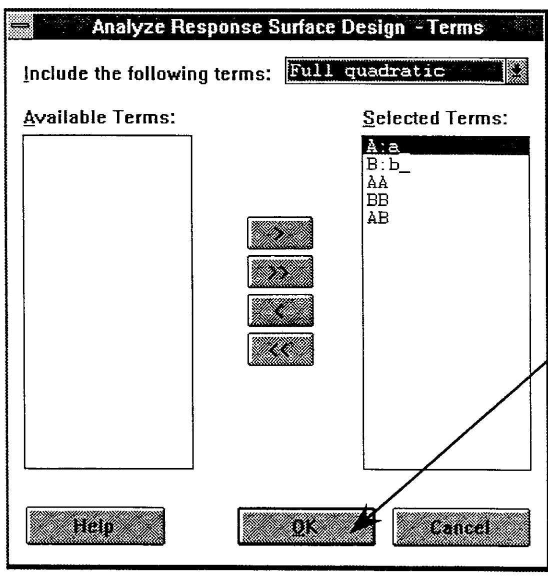


Click on 'Fit Model. . .'

Fill in the main dialog box as shown:



Select the 'Terms' button from the main dialog box to specify the type of model ('full quadratic') and the factors to be included in the model (all "X"s, squared "X"s, and interactions):



Fill in the dialog box as shown

Click 'OK' to return to main dialog box

Session Window output:

Response Surface Regression

Estimated Regression Coefficients for twists

Term	Coef	StDev	T	P
Constant	48.312	2.15998	22.367	0.000
a_	7.775	0.95122	8.174	0.000
b_	-1.655	0.19024	-8.699	0.000
a_*a_	-1.062	1.06350	-0.999	0.341
b_*b_	-0.057	0.04254	-1.352	0.206
a_*b_	-0.054	0.17016	-0.317	0.758

statistically significant (p < 0.05)

NOT statistically significant (p > 0.05)

S = 4.254 R-Sq = 93.6% R-Sq(adj) = 90.3%

Analysis of Variance for twists

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	5	2631.47	2631.47	526.29	29.08	0.000
Linear	2	2578.53	2578.53	1289.26	71.24	0.000
Square	2	51.12	51.12	25.56	1.41	0.288
Interaction	1	1.82	1.82	1.82	0.10	0.758
Residual Error	10	180.97	180.97	18.10		
Total	15	2812.44				

only 'Linear' is statistically significant!

Unusual Observations for twists

Obs	twists	Fit	StDev Fit	Residual	St Resid
3	69.000	61.315	2.462	7.685	2.22R

Evaluation of Model:

- "a_" and "b_" are statistically significant factors
- From the ANOVA table, only the Linear model is statistically significant; Square and Interaction don't make a difference in the "Y", so a linear model is best for this data
- R^2 and R^2adj are over 90%, which indicates a potentially good fit
- 's' (the standard deviation of the error term) is 4.254. This is the "sigma" of the unexplained variation - noise not included in the model. Decide if more "X"s need to be included by looking at +/- 6s (~25, in this case). Can you live with +/- 25 twists variation, even if 'a_' and 'b_' are controlled perfectly?

Start with a quadratic model and re-fit to a simpler model if possible

Re-Fitting the Model With a Simpler Equation

Using the coefficients presented in the table from the previous page, our **quadratic** model for Twists would be:

$$Y = 48.312 + 7.775 (a\_) - 1.655 (b\_) - 1.062 (a\_ * a\_) - 0.057 (b\_ * b\_) - 0.054 (a\_ * b\_)$$

Then, if we UN-center the "X"s (in order to make the equation useful from a practical standpoint):

$$Y = 48.312 + 7.775 (A - 2.5) - 1.655 (B - 12.5) - 1.062 (A - 2.5)^2 - 0.057 (B - 12.5)^2 - 0.054 (A - 2.5)(B - 12.5)$$

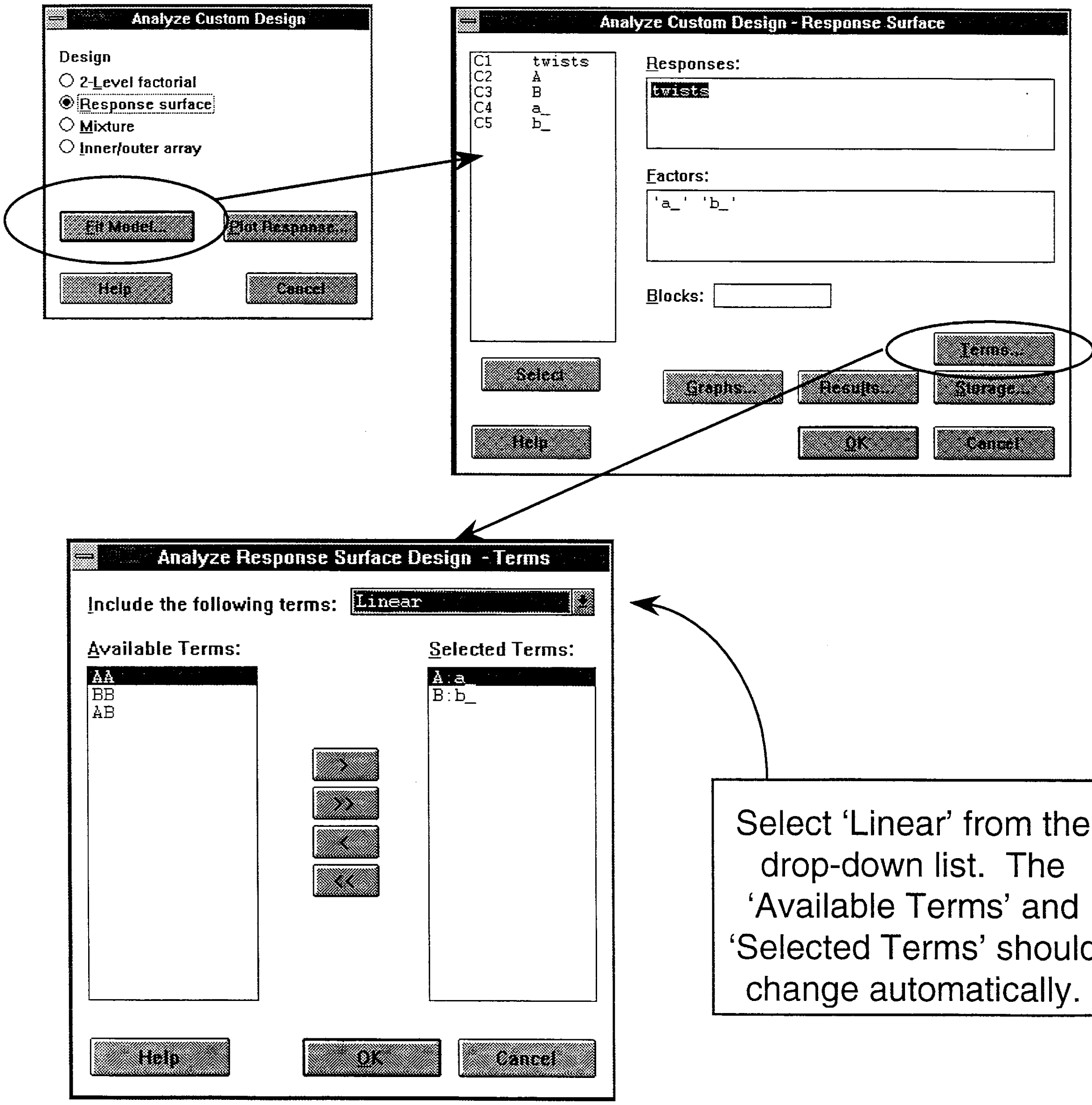
With all of the quadratic and interaction terms, this equation is pretty complicated!

The data told us that only the Linear terms were significant, and the Residuals graphs told us that the Quadratic model may not be the best fit. Let's re-do the analysis, selecting "Linear" instead of "Quadratic" . . . .

Re-Fitting the Model

Use 'Ctrl-e' to return to the 'Analyze Custom Design' dialog box. The only aspect we will change is in the 'Terms' dialog box - everything else can stay the same:

Retain the selections in the first two dialog boxes:



Select 'Linear' from the drop-down list. The 'Available Terms' and 'Selected Terms' should change automatically.

Click 'OK' twice

Select Residual plots from 'Graph' and 'Storage' boxes

Session Window Output

Response Surface Regression

Both "X" variables and the Linear model are statistically significant

($p < 0.05$)

Estimated Regression Coefficients for twists

Term	Coef	StDev	T	P
Constant	45.187	1.0605	42.611	0.000
a_	7.775	0.9485	8.197	0.000
b_	-1.655	0.1897	-8.724	0.000

S = 4.242 R-Sq = 91.7% R-Sq(adj) = 90.4%

Analysis of Variance for twists

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	2	2578.53	2578.53	1289.26	71.65	0.000
Linear	2	2578.53	2578.53	1289.26	71.65	0.000
Residual Error	13	233.91	233.91	17.99		
Total	15	2812.44				

- Both R^2 values are similar for the Linear and Quadratic models, indicating that the Linear model still provides a good fit.

	Quadratic	Linear
R-Squared	93.6	91.7
R-Squared Adjusted	90.3	90.4

- The spread of the error terms is similar for the both models: 4.242 (Linear) and 4.254 (Quadratic)

Linear Equation:

Twists = 45.187 + 7.775 (a_) - 1.655 (b_) ***Centered Data***

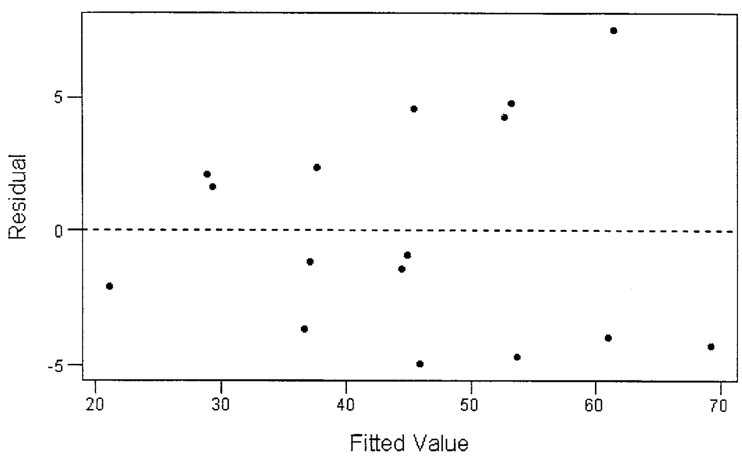
OR

Twists = 45.187 + 7.775 (A - 2.5) - 1.655 (B - 12.5) ***Uncentered Data***

Next Step: *Review Residual Plots*

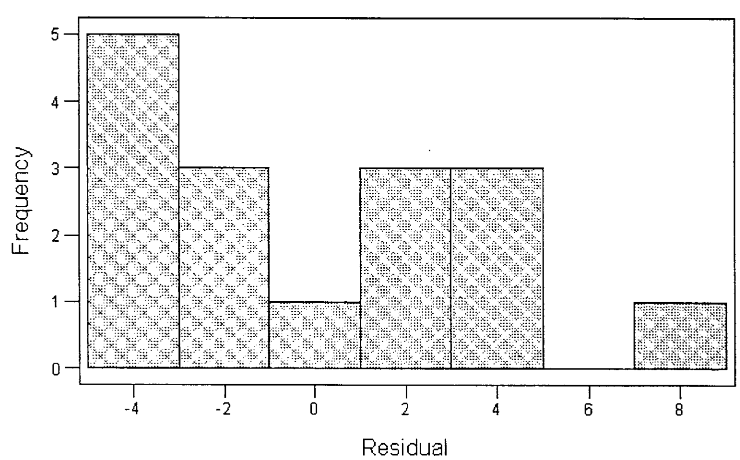
Use Residual Plots to Help Evaluate the Linear Model

Residuals Versus the Fitted Values
(response is twists)



There is a possible increase in variation with higher fitted values.

Histogram of the Residuals
(response is Twists)



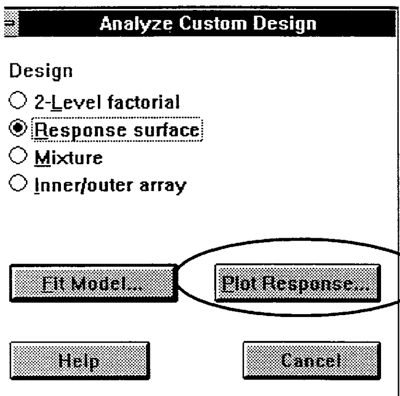
Residuals don't look very normal (remember we only had 15 data points)

<u>Interpretation for the Linear model</u>: All assumptions may not be strictly met. However, this model provides a good approximation to the data and will likely be beneficial in practice. If a better fit is necessary, the data may require transformation - possibly use the log or reciprocal of "Y".

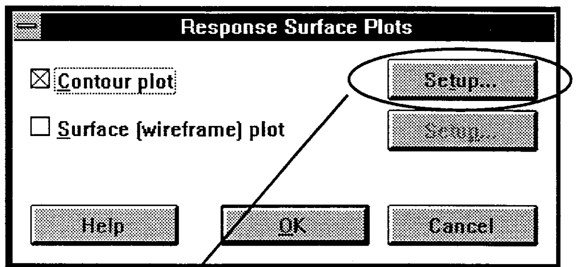
LAST STEP: Look at the 'Contour Plot'. . .

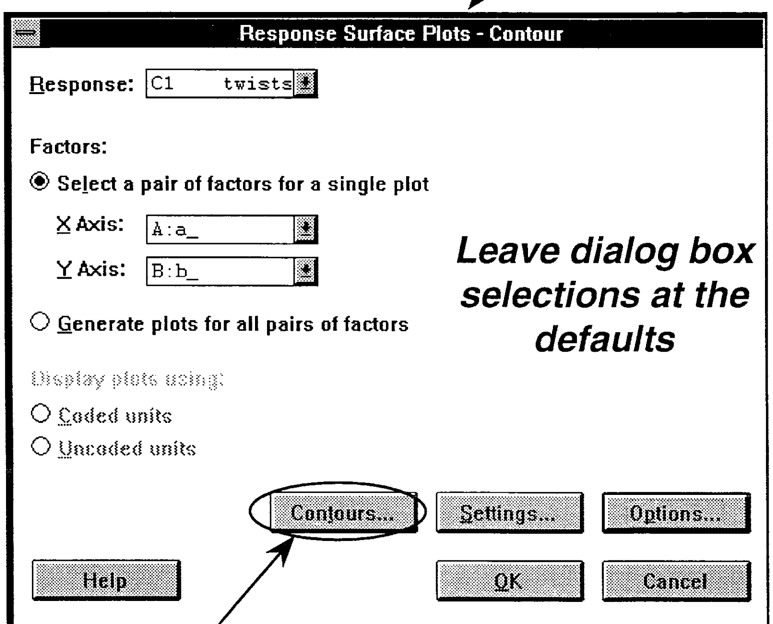
Look at Contour Plots to Find Optimal Operating Conditions

Use 'Ctrl-e' to return to the opening dialog box. Select 'Plot Response' . . .



Click 'Contour Plot', then 'Setup' . . .



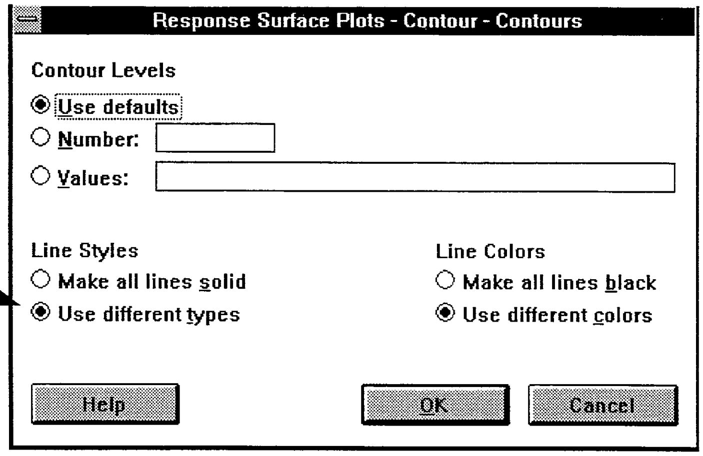


Leave dialog box selections at the defaults

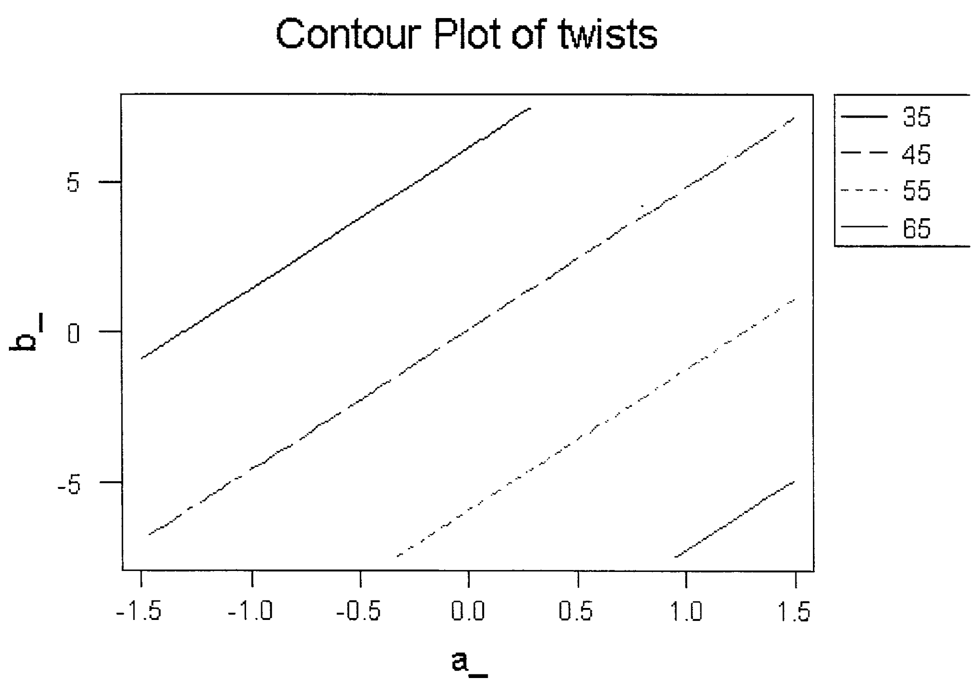
Click on 'Contours . . .' to set up the graphical output . . .

Contour Plot

Click 'Use different types' to create dashed lines in the Contour Plot



Click 'OK' three times!



Lines of constant response for "Twists"

Interpretation: to maximize the number of Twists, move towards the lower right-hand corner of the Contour Plot (Twists = 65). Read off potential "a_" and "b_" values that will provide Twists = 65.

(remember to un-center the "X" values to obtain true process settings!)

One more tool: 'Stepwise' Regression

Regression can be a valuable tool in the screening process to narrow down a large number of "X"s to the Potential Vital Few. This can even be done using Baseline data, but <u>be careful</u>:

<u>CAUTION</u>!!

<u>NEVER</u> draw conclusions about which "X"s are the Vital Few without first performing a DOE to confirm that these really are the "X"s that control the process

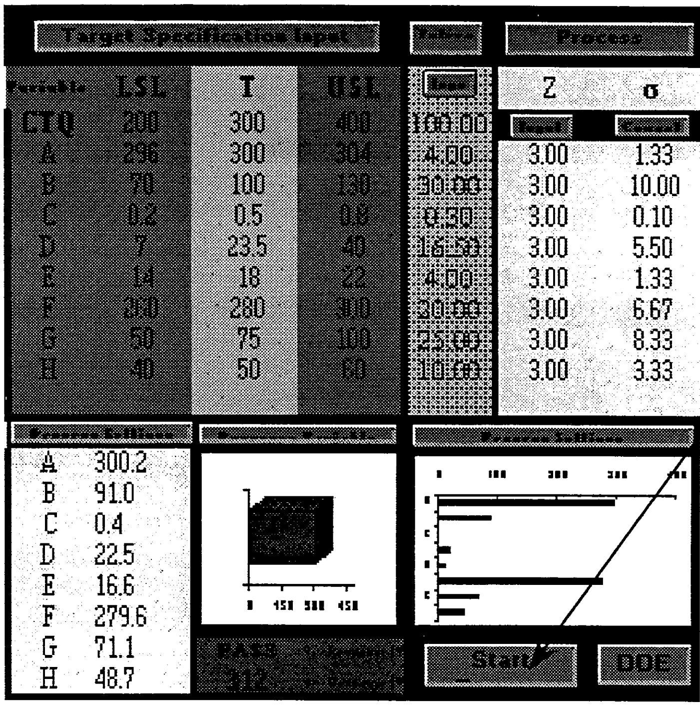
In 'Stepwise' Regression, "X"s are progressively added to the model based upon their influence on the response ("Y"). The first "X" used by Minitab is the one with the largest influence.

Let's practice Stepwise Regression using a computer-generated factory known as the 'Factory Simulator'. This factory makes 'widgets'. The CTQ ("Y") is <u>widget length</u>.

The Factory generates a <u>continuous</u> "Y" response using 8 <u>continuous</u> "X" variables (labeled A through H)

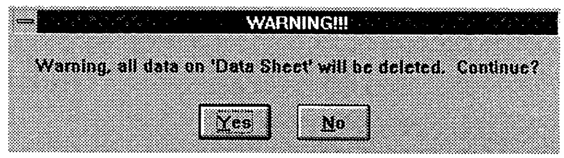
The Factory Simulator is an Excel file: **<u>FACTSIM.xls</u>**

Running the Factory Simulator

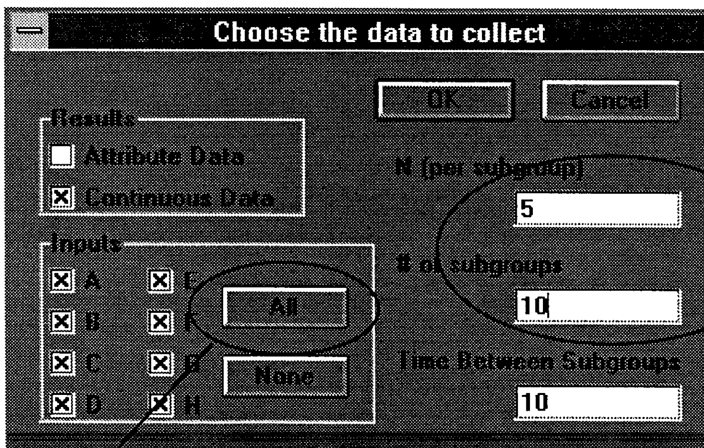


The Factory Simulator makes 'widgets' (the red boxes shown at bottom center)

Click the 'Start' button to start up the Factory



Click 'Yes'



Choose the number of samples in the subgroup

AND

Choose the number of subgroups

Click 'All' to select all 8 "X" variables

Follow the step-by-step procedure on the next page to use 'Stepwise' Regression to screen for the Potential Vital Few "X"s

USING STEPWISE REGRESSION TO SCREEN FOR POTENTIAL "VITAL FEW" X's (cont'd)

The basic regression formula is:

$$y = m_1x_1 + m_2x_2 + m_3x_3 + \ldots + m_nx_n + b$$

For our factory simulator, the formula would be:

$$\text{Factory Output} = C_1A + C_2B + C_3C + C_4D + C_5E + C_6F + C_7G + C_8H + b$$

where $C_1, C_2 \ldots C_8$ are coefficients, and "b" is the constant (actually the y-intercept of the line)

Steps in the process:

1. Run the data off of the factory simulator (make sure you have a large enough sample with appropriate rational subgroups)

2. Copy the output data to Minitab, being sure to delete the blank lines FIRST.

3. Graph the output data using **GRAPH>MATRIX PLOT.** This will allow you to compare the output (Y) to each X variable (A, B, C, D, etc.).

In the dialog box, select the output (response column) and 4 of the input variable columns. Click **OK**. Do this again to plot the output versus the other 4 input variables. These graphs will show trends. Do any of these Y vs. X plots look linear? Are any curved?

4. Select **STAT>REGRESSION >STEPWISE**

"Response": Y (your output column - C1)

"Predictors": A through H (columns 2 through 9)

Click the "Options" button and enter "1" in the box where it says

"Take _____ steps between pauses".

Click "OK" twice.

Minitab will prioritize the influencing "X" variables and run the first regression step on the one with the greatest influence. Then Minitab will ask if you want to run more. Type "yes", and hit return. Continue doing this until Minitab won't calculate any more. At that point, you will have all of your ***POTENTIAL*** "Vital Few" !

IMPORTANT NOTE: On the last step Minitab calculates, check the R^2 value. Only **you** can decide if the fit is good enough to use this model!

Key Concepts: Tab 8 Multiple Regression

- Always graph your data first!

- If a scatterplot of the data shows a potential curved relationship between "Y" and the "X"s (or if you aren't sure and need a starting point for the analysis), fit the data with a quadratic model - 'Response Surface'.

- To separate the effects of "X" from the effects of "X"^{2}, 'center' the data by subtracting the column average from each "X" value (this ensures orthogonality) .
  Use centered values in all analyses, transform back to un-centered values when complete.

- Look at p-values, R^2 and R^2adj, and s values for initial evaluation of the model; use residual plots to check error terms.

- Use the Contour Plot to find the combination of "X" values needed to generate the desired "Y".

- Use Stepwise regression to progressively add "X"s to the model based upon their influence on "Y".

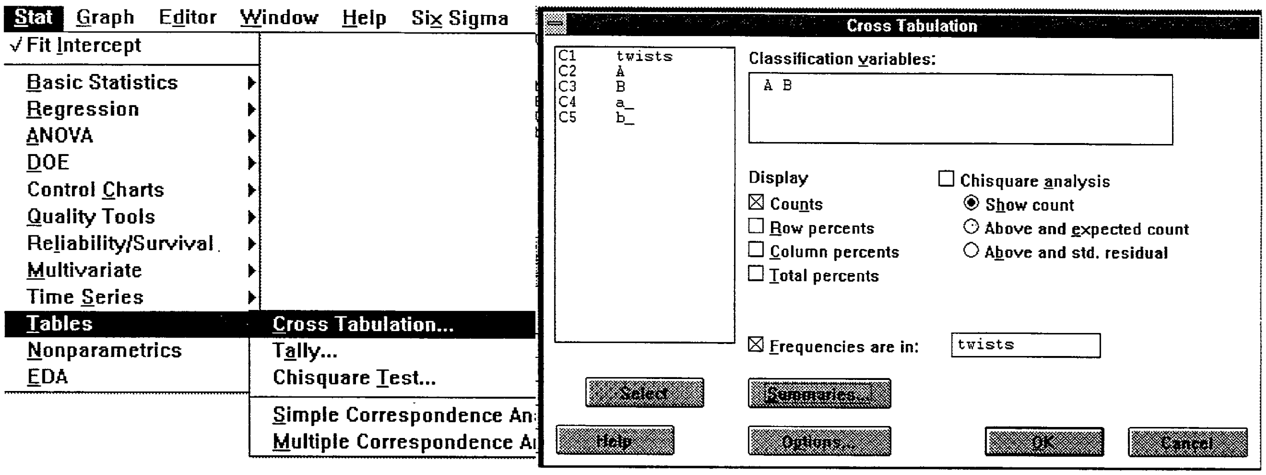
Appendix

Using Cross-Tabulation to Look for Trends in the Data:

(Using the "Twists" data from the beginning of this tab)

How the data is arranged can provide clues to the significant "X"s . . . .

Stat>Tables>Cross Tabulation Select 'Display' **Counts**



Rows: A Columns: B

	5	10	15	20	All
1	41	40	31	19	131
2	49	50	36	31	166
3	69	58	44	33	204
4	65	57	57	43	222
All	224	205	168	126	723

Cell Contents --
Count

Look for highest number of Twists - what combination of 'A' and 'B' alloy maximizes the twists?

Based on this table, it appears that higher percentages of alloy 'A' and lower percentages of alloy 'B' maximize the number of Twists. Use this information to double-check the results of the regression analysis . . .

Multiple Regression - the mathematical viewpoint

Multiple regression is used when there are multiple independent variables (X's) and one response (Y).
The model may include quadratic terms, but the estimated coefficients (b's) are linear.
The model is linear in the parameters (b's).

$$Y_i = a + b_1 * X_{1i} + \ldots + b_k * X_{ki} + e_i$$

[You may choose to ignore the following]

$$(y) = [X] \quad (b) + (e)$$

( ) denotes a vector.
[ ] denotes a matrix

$$(y) = \underbrace{\begin{pmatrix} Y_1 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{pmatrix}}_{\text{response}} \qquad [X] = \underbrace{\begin{bmatrix} 1 & X_{11} \ldots X_{k1} \\ \cdot & \cdot \quad \cdot \\ \cdot & \cdot \quad \cdot \\ \cdot & \cdot \quad \cdot \\ 1 & X_{1n} \ldots X_{kn} \end{bmatrix}}_{\text{Independent variables}}$$

$$(b) = \underbrace{\begin{pmatrix} a \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_k \end{pmatrix}}_{\text{estimates of parameters}} \qquad (e) = \underbrace{\begin{pmatrix} e_1 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{pmatrix}}_{\text{errors}}$$

Least squares solution (minimize sum of e_i's) is: $(b) = (X'X)^{-1} X' y$

Excel Solver

What is it?

A function in Excel that allows you to solve equations (determine "X" values) for specific values of "Y", or find a minimum or maximum (for quadratic relationships).

Where is it used?

After a model equation has been generated through regression analysis, DOEs, etc.

How does it work?

It solves equations (or sets of equations) using the technique of partial differentiation. It's better not to go there - let Excel do the work for you!

What are the limitations of Solver?

The biggest concern: be sure you are solving the equation within the boundaries that you tested for the "X"s. In other words, DON'T EXTRAPOLATE!

Using Excel Solver to solve the regression equation

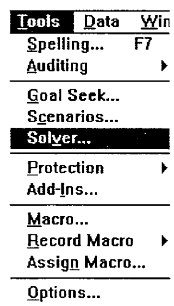
Step 1: Open Excel to a blank spreadsheet. Enter the uncentered linear equation into the spreadsheet in cell B4, as shown. Hit the 'Enter' key when complete.

B4 = 45.187+7.775*(C3-2.5)-1.655*(D3-12.5)

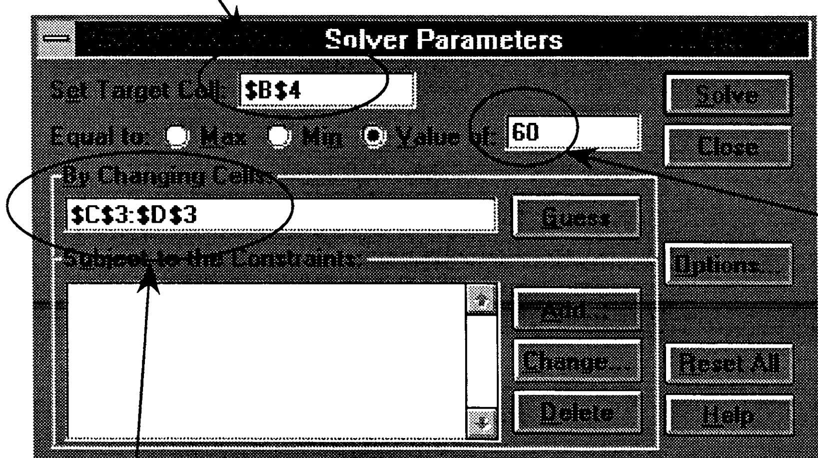
	A	B	C	D	E	F	G
1							
2		Y	A	B			
3							
4		= 45.187+7.775*(C3-2.5)-1.655*(D3-12.5)					
5							
6							
7							

NOTE: cells C3 and D3 are 'reference' cells for the values of A and B. When Solver solves the equation, the values of the coefficients will be placed in these two cells.

Step 2: Open Excel Solver: **Tools>Solver**



Click on B4 (the cell containing the formula)



Tell Solver to solve the equation for a value of **Y = 60 twists**

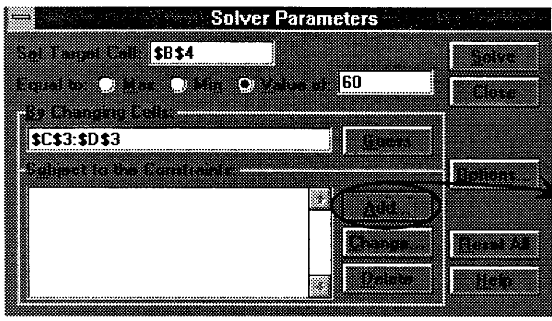
Highlight the cells where the answers will be displayed: C3-D3

Continued . . .

Using Excel Solver to solve the regression equation (cont'd)

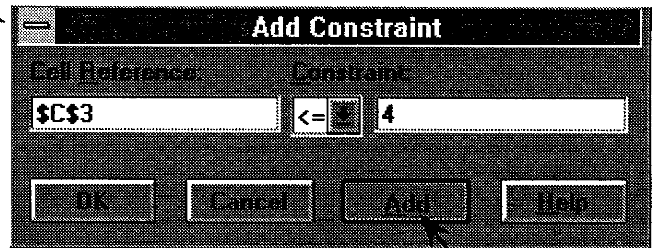
Step 3: Enter the limitations on the "X"s (the test ranges)

<u>REMEMBER</u>: you CANNOT extrapolate beyond the test region!

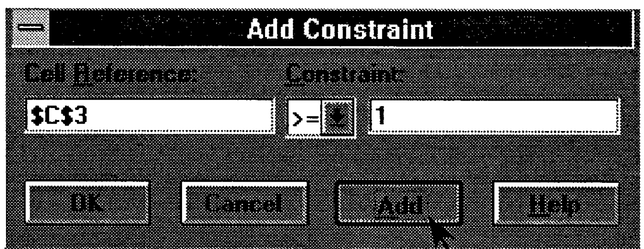


Enter the range for "A": 1 - 4

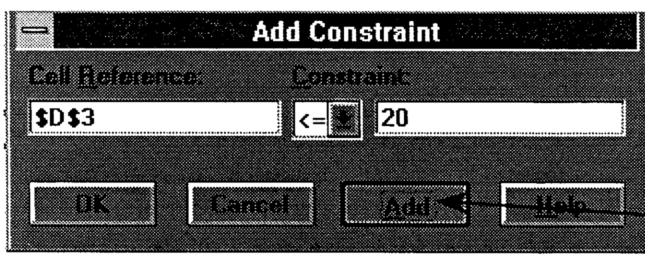
Enter the range for "B": 5 - 20



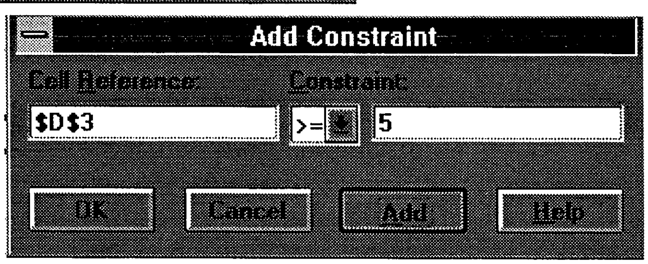
Click 'Add'



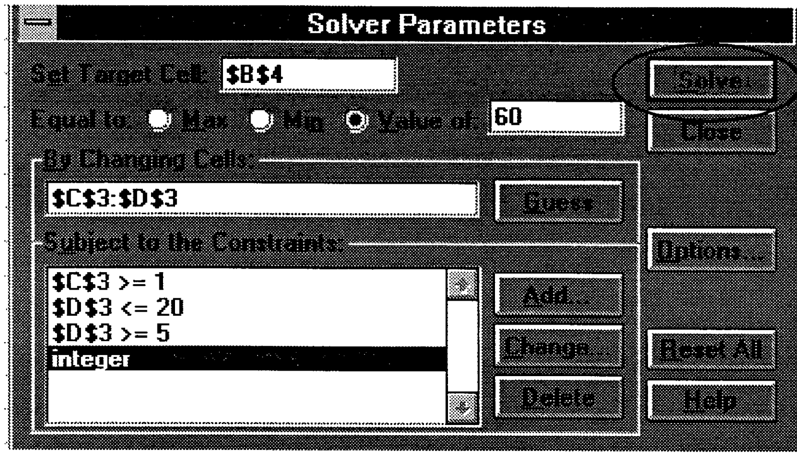
Click 'Add'



Click 'Add'

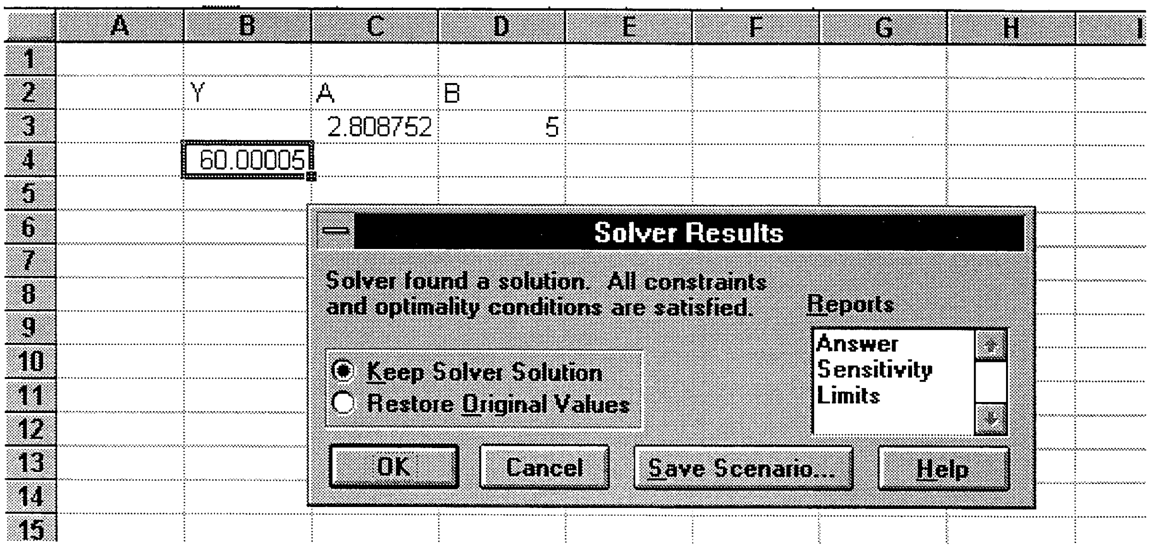


Click 'Add', then Click 'Cancel'



Click 'Solve' to generate the answers for "A" and "B"

Here is one possible setting for A and B:



Excel Solver can provide specific values of the "X"s to generate either a particular value for "Y", or to find a minimum or maximum value of "Y" (for quadratic equations only)