SAS INSTITUTE INC.
FALL 1996
ISSUE 3

JMPer Cable®

NEWSLETTER FOR JMP® USERS

DEEP INTERACTIVITY

John Sall
Senior Vice President
SAS Institute Inc.

The name JMP was chosen partly to convey the idea of a lively product that jumped responsively to the user's wishes. This article examines why the fine-grained progressive nature of JMP's interactivity makes it feel more lively than other statistical products.

The reason that other software products don't feel as lively as JMP is that they really aren't interactive. They give the illusion of being interactive by having dialogs that you fill out interactively, and output that you can view and customize interactively. But the statistical heart of the analysis lives only a brief life and is not interactive at all. Under the skin of these products is an old patched-up batch program that has been front-ended with a mask of dialogs and back-ended with a word processor or a graphics editor. The dialogs are nice in many ways. They eliminate much of the learning and remembering burden of a command language system. So what's wrong?

The key fault of these packages is that they force you to front-load all decisions before you see any results. You fill out pages of dialog that offer you dozens of features, diagnostics, reports, and graphs, and you have to decide right then what you will want. Do you want a plot of the residuals? You'd better say yes now because if you say no and change your mind, you will have to go back and rerun the whole analysis.

The problem with front-loading decisions is that the decisions have nothing to do with the results. There is no interactivity. There are no feedback cycles. Instead of exploring the data firsthand, you are an absentee bureaucrat handing it a book of filled-out forms. That may work fine when the goal is only to do a prescribed analysis. But it makes it harder to explore, and it makes it less likely that you will discover interesting and valuable phenomena. You spend more time filling out forms, and less time looking at the results. The dialogs demand your attention on their schedule, rather than provide interactive features that are at your service when you want them.

The lack of interactivity leads to a number of other problems in the design of statistical products. There are more analyses and each analysis is more specialized. For example, Instead of presenting a unified interactive platform for exploring a distribution, there are separate commands for descriptive statistics, histograms, box plots, and distributional tests. This gives more interactivity in a sense because you have to interact with the product each time to ask for each element, and you see results before doing the next step. But this interactivity is a burden, not a boon. You are throwing away the context of the analysis you have established. You have to go back to the top of the command tree, find the new command, and go through its dialogs, selecting again the variable of interest. The results come out in a separate window, and you end up having to examine and close many windows instead of one.

Then the system grows to adapt to the problems it creates. To compensate for the burden of doing many little commands, you are offered scripting or macros or templates to automate them. To compensate for having separated output, you are offered a window in which to paste results together. To compensate for having too many commands offered at the top, you are offered another system of dialogs to select the right subcommands. ➡

(continued on page 2)

> Comment from an anonymous graduate statistics student:
>
> *"I took a statistics class as an under graduate, but I never really put the regression tools in my head—they fell out after the final exam. I think with JMP I'll be able to take it with me this time."*

IN THIS ISSUE

THE CASE OF THE MISSING FORMULAS

Occasionally when doing table manipulations such as subsetting, sorting, splitting or stacking columns, or joining tables you might be surprised (dismayed) to find that the formula you worked hard to build has disappeared. Usually this isn't a problem because the Tables menu commands produce new data tables, and formulas remain intact in the original table. You can cut and paste to retrieve them if necessary.

Further, the circumstances that lead to a disappearing formula are logical. The following list tells you when formulas are saved or lost by Tables menu commands:

- **Tables→Subset**
  When you subset selected columns and include a formula column but don't include all its column elements, that formula is dropped from the new table, but the computed column values remain.

  Rationale: The formula in the subset table would have empty elements, which could then result in missing column values.

- **Tables→Sort**
  Formulas in any Sort key variables you specify are not preserved.

  Rationale: There could be a conflict between the sorted order of a variable and the order produced by its computed value. This is often the case if a computed value depends on a function of *i*, the row number.

  However, you do have the option on the Sort dialog to 'sort in place.' That is, you can have the sort operation replace the original table with the sorted table. In that case the formula is retained. But be aware that the formula does not automatically re-evaluate. You make the choice to either recompute the column with its calculator or to remove the formula from the column through the Column Info dialog.

- **Tables→Stack Columns**
  If any of the columns to be stacked contain formulas, those formulas are not kept. Also, if a column has a formula that has as elements any of the columns to be stacked, its formula is not kept.

  Rationale: 1) If you stack computed columns, their formulas might not all be the same, and there is no way to choose which one to keep. 2) Any formula that depends on *i*, the current row number, or *n*, the total number of rows, could give unexpected results and is dropped. 3) Because the columns in the stack list will no longer have their individual column names, any formula in other columns that refers to them is dropped.

- **Tables→Split Columns**
  If the column to be split has a formula, that formula is not passed to the resulting split columns. Also, if a column has a formula that uses any of the columns to be split as an element, its formula is not kept in the new table. In addition, any formula that depends on a column used as a Col ID is also dropped.

  Rationale: The reasons for dropping formulas when splitting columns are the same as those for stacking columns. Also, if splitting columns uses an ID column, that column does not exist in the new table; therefore, any formula dependent on it would have missing elements.

- **Tables→Join**
  If a column formula depends on another column element that is dropped (not kept) from the joined table, the formula is not kept but the computed values remain. Joining with matching variables when both matching columns have a formula, and keeping non matches keeps both formulas but you have to reevaluate the columns. Similarly, if you join by matching columns that have formulas and drop one of those columns, the other retains its formula. However, you should reevauate to be sure the results are correct. In general, make the choice to either recompute the column with its calculator or to remove the formula using the Column Info dialog.

  Rationale: Formulas that depend on columns that are dropped in the joined table would have empty elements, which could then result in missing column values.

- **Tables→Transpose**
  . . .left as an exercise for the reader.

(continued from page 1)

To compensate for losing the context of the analysis, you are offered options to replay the forms you last submitted so that you can change the options.

Contrast this with JMP. Analyses are interactive and progressive. At the top, there are only a few general platforms; the details unfold as the context develops in the course of an analysis. You can decide what to do at any time, not just when you launch the analysis. All the results appear grouped in one window with a highly interactive surface. And under the interactive surface sit interactive statistical engines that keep the context and the internal calculations ready to do the next subanalysis.

With many products, you feel like you are moving from bus station to bus station, looking up schedules, filling out ticket purchase forms, and staying on the routes. With JMP, you feel you are driving, and can explore the data at will. Bon Voyage.

MORE POWER!

by Duane Hayes
SAS Institute Inc.

Everyone wants power. Be it political, muscular, mental or statistical power. From Captain Kirk of the Starship Enterprise to Mr. Felton in your company's statistics department—everyone wants power. With this in mind, JMP software offers an easy way to calculate at least one type of power… statistical power.

Power, as I am referring to it here, is the probability of getting a significant statistical test result for a given alpha level (the probability of obtaining a Type I error). To calculate power, you need to know the alpha level of the statistical test, an estimate of the standard error of the test (σ), an effect size (δ), which is the size of the difference you want to detect, and the sample size.

To increase the power of a statistical test, you have several options:

- increase the alpha level
- decrease the standard error of the test
- increase the effect size
- increase the sample size.

In many cases, the first three options are neither suitable nor possible, so the only choice is to increase the sample size. But, this raises the question: "How many samples are needed?"

JMP to the Rescue!

For any least squares model, the model fitting platform (Fit Model command) in JMP gives a power details option for the whole model and for each of its effects. JMP calculates initial estimates for the power parameters and displays them in an editable dialog. You can then enter the values (or ranges of values), and JMP calculates power and sample size for all combinations of these values.

Let's say you plan an experiment to compare weight gain in mice using four food additives. Suppose you want to detect a minimum difference in weight gain of ten grams between any two food additive groups, with a probability of making a Type I error of 0.05, (that is, α=0.05). From knowledge of earlier experiments, you have an estimate of the standard error as σ= 3 and you want the power of the test to be at least 0.90. How many replicates are needed?

With these parameter values, you can use JMP to do a *prospective power analysis*. This kind of analysis is a way to get JMP to estimate power for a range of sample sizes and minimum differences (deltas) before actually conducting an experiment.

First, create a dummy data table similar to the one shown in **Figure A**. The values for the mean weight gain are not important because you are going to supply your own parameter values.

Next, choose Analyze→Fit Model and complete the Specify Model dialog that appears. Specify, GAIN(Y) as Y and ADDITIVE as the only effect in the model.

Figure A WTGAIN.JMP Table

3 Cols / 5 Rows	Additive	Gain(Y)
1	A	1
2	B	1
3	C	1
4	D	1

Click the Defer Plots check box because plots are of no interest here, and then click Run Model.

When the analysis results appear, select the pop-up menu next to the Whole Model Test table and choose Power Details to see the Power Details dialog shown to the left in **Figure B**. Initially, it contains estimates given by the analysis.

Click the Solve for Power check box and enter the following values for the parameters as shown to the right in **Figure B**. Enter .05 for Alpha, 3 for Sigma, 3.5355 for Delta, and the range From 4 To 40 By 4 for Number.

When you click Done, JMP calculates power estimates for all combination of values in the dialog and gives the results shown in **Figure C**. You can see that a total sample size of 16 (3 replicates for each food additive) ensures a power of 0.90.

What Is Delta?

To complete the Power Details dialog you needed to enter a value for Delta. The key here is to get JMP's delta in terms of the minimum difference between means you want to detect. Montgomery (1984) and a

Figure B Power Details Dialog

Power Details Dialog

Whole Model
Click and Enter 1, 2 or a sequence of values for each:

	Alpha	Sigma	Delta	Number
From:	0.050	0	0	4
To:	•	•	•	•
By	•	•	•	•

- [ ] Solve for Power
- [ ] Solve for Least Significant Number
- [ ] Solve for Least Significant Value
- [ ] Adjusted Power and Confidence Interval

Done | Cancel | Help

Calculations will be done on all combinations

Power Details Dialog

Whole Model
Click and Enter 1, 2 or a sequence of values for each:

	Alpha	Sigma	Delta	Number
From:	0.050	3	3.5355	4
To:	•	•	•	40
By	•	•	•	4

- [x] Solve for Power
- [ ] Solve for Least Significant Number
- [ ] Solve for Least Significant Value
- [ ] Adjusted Power and Confidence Interval

Done | Cancel | Help

Calculations will be done on all combinations

Figure C Power Details Analysis Results

Power Details

Test Whole Model

Power

Alpha	Sigma	Delta	Number	Power
0.0500	3	3.5355	4	•
0.0500	3	3.5355	8	0.3953
0.0500	3	3.5355	12	0.7678
0.0500	3	3.5355	16	0.9317
0.0500	3	3.5355	20	0.9831
0.0500	3	3.5355	24	0.9963
0.0500	3	3.5355	28	0.9993
0.0500	3	3.5355	32	0.9999
0.0500	3	3.5355	36	1.0000
0.0500	3	3.5355	40	1.0000

little algebra provide the answer. First note the following terms. Let

D = the minimum difference to be detected

a = the number of treatment levels

n = the number of replicates

σ^2 = estimated standard error.

If $\mu_1, \mu_2, ..., \mu_a$ are the specified treatment means, then define

$\tau_i = (\mu_i - \text{mean}(\mu))$,

where

$$\text{mean}(\mu) = \frac{\Sigma\mu_i}{a}$$

is the average of the individual treatment means.

It follows

$$\frac{n\Sigma\tau_i^2}{a\sigma^2} = \frac{nD^2}{2a\sigma^2}$$

This leads to

$$D^2 = 2\Sigma\tau_i^2$$

JMP's definition of Delta^2 is

$$\delta^2 = \frac{\Sigma(\mu_i - \text{mean}(\mu))^2}{a} = \frac{\Sigma(\tau_i^2)}{a}$$

Substituting back, it can be shown that

$$D^2 = 2a\delta^2$$

So JMP's $\delta = \frac{D}{\sqrt{2a}}$

In this example the delta to enter is

10 ÷ sqrt(2*4) = 3.5355

To Summarize

Once you understand how to do a prospective power analysis and how to enter the effect size you want, power and sample size determination will no longer be confusing, and you will always be able to see how to obtain the power you desire (at least statistically).

Montgomery, Douglas C. (1984), *Design and Analysis of Experiments*, New York: John Wiley and Sons.

JMP Statistics and Graphics Guide, Version 3.1, (1995), SAS Institute, Inc.

Calculator Corner

by *Michael Hecht*
SAS Institute Inc.

Currently, JMP has a color palette of 12 colors (excluding black), called hues. There are 5 shades of each hue. For plots it is sometimes useful to create a rowstate column that assigns colors to data table rows based on a response variable. This is easily done with a formula that uses functions from the Row State and Statistical calculator functions.

Hues

To uniformly assign one of 12 hues to a range of numbers (call it Y), you first use the quantile function to scale and divide Y into 12 equal intervals.

To find the intervals of Y, first subtract the minimum, $\text{quantile}_0\,Y$, from each response Y. Then divide that difference by the range of the response, $\text{quantile}_1\,Y - \text{quantile}_0\,Y$ ($\text{quantile}_1\,Y$ is the maximum of Y). So the formula is:

$$\frac{Y - \text{quantile}_0\,Y}{\text{quantile}_1\,Y - \text{quantile}_0\,Y}$$

You then use the Hue function to assign colors. However, the Hue function interprets numbers from 1 to 12 (not fractions from 0 to 1). To modify the interval values for the hue function, multiply them by 12. The complete formula for assigning 12 unique colors becomes:

$$\text{hue}\left(\frac{Y - \text{quantile}_0\,Y}{\text{quantile}_1\,Y - \text{quantile}_0\,Y} \cdot 12\right)$$

To try this yourself, create a new column. On the New Column dialog select Row State as its data type and Formula as its Data Source. Then enter the formula above into its calculator. Use any response you want as Y.

Shades

The Shade function interprets numbers from –2 to 2, giving 5 shades. You use a formula similar to the one shown for the Hue function to assign one of 5 shades.

Because the internal values that define shades range from -2 to 2, (instead of 1 to 5), you have to subtract 2 from the shade formula:

$$\text{shade}\left(\frac{Y - \text{quantile}_0\,Y}{\text{quantile}_1\,Y - \text{quantile}_0\,Y} \cdot 5 - 2\right)$$

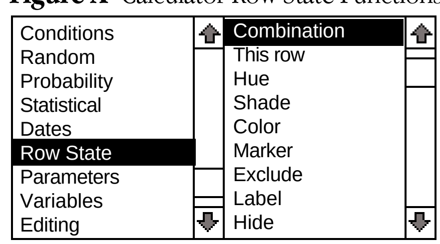
Shortcut: To automatically generate these Hue and Shade formulas, select your Y variable and then OPTION-click (Mac) or CNTL-click (Windows) on Hue or Shade

Hues and Shades Together

To assign both hue and shade requires extending the formula to include both the Hue and the Shade functions. To do this, you create a row state column as before and use its calculator.

Select Combination from the list of Row State functions as shown in **Figure A**.

Figure A Calculator Row State Functions



The Combination function sets up an empty function like the one shown « □,□ » shown to the right. You fill in the two arguments with the Hue and Shade formulas shown previously, giving the combination of row state functions shown in **Figure B**.

Note: Use the calculator's insert button to add arguments to the Combination function for applying more than two row states (Hide, Label, Marker or Select).

Figure B Combination Function Showing Hue and Shade

$$«\ \text{hue}\left(\frac{Y - \text{quantile}_0\,Y}{\text{quantile}_1\,Y - \text{quantile}_0\,Y} \cdot 12\right),\ \text{shade}\left(\frac{Y - \text{quantile}_0\,Y}{\text{quantile}_1\,Y - \text{quantile}_0\,Y} \cdot 5 - 2\right)\ »$$

DOSE-RESPONSE CURVES AND INVERSE PREDICTION

by Annie Dudley
SAS Institute Inc.

In drug studies, logistic models and curves are often used to predict the rate of response for different doses of a particular drug. The logistic cumulative distribution function lends itself well to these types of studies because it is S-shaped and gradually tapers off toward the extremes, while having a steep change in the center.

To estimate the dose (or regressor, X) at a particular probability (p) of an event, choose a value for p and solve the equation

$$\hat{X} = \frac{\ln\left(\frac{p}{1-p}\right) - b_0}{b_1}$$

where b_0 and b_1 are parameter estimates. The dose that yields a 50% lethal response (LD50) or 50% cure (ED50) is often used by the researcher as a preliminary statistic for comparing the potency of different drugs.

Figure A shows example dose-response data for a logistic regression analysis.

Figure A Streptococci Data

10 Rows / 4 Cols	Dose	Response	Count	ln(dose)
1	0.125	Cured	0	-2.07944
2	0.125	Died	11	-2.07944
3	0.25	Cured	3	-1.38629
4	0.25	Died	9	-1.38629
5	0.5	Cured	8	-0.69315
6	0.5	Died	4	-0.69315
7	1	Cured	11	0
8	1	Died	1	0
9	4	Cured	7	1.386294
10	4	Died	0	1.386294

The data are from an experiment where rabbits were injected with different amounts of penicillin to test the extent of protection from potentially lethal injections of beta-hemolytic streptococci. The response is whether the rabbits were cured or died.

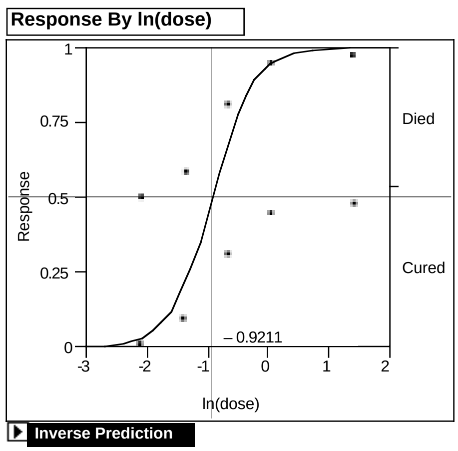
Note that the data table in **Figure A** is summarized to show the frequency (Count) for each combination of dose and response.

Often the natural log of dose is the preferred response for a dose-response analysis. You can use the JMP calculator to create a column like ln(dose) in this example.

Use these steps to do a dose-response analysis and show the response curve:

- Be sure the response is either nominal or ordinal (N or O). If the response is coded numeric (0 or 1), use the popup menu at the top of the column and change the modeling type to N or O.
- Choose Analyze→Fit Y by X and select Response as Y and ln(dose) as X. The example data are summarized so Count is the Freq variable. When you click Done you see the ln(dose) response curve in **Figure B**.

Figure B ln(dose) by Response With Crosshair



The analysis includes the supporting statistical tables shown in **Figure C.** The test of the ln(dose) parameter estimate shows that increasing penicillin dose significantly increased probability of survival. Together with the significant whole model chi-square test, the logistic curve appears to be a good fit for the data.

Figure C Logistic Regression for Response by ln(dose)

Converged by Gradient

Whole-Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	18.970251	1	37.9405	<.0001
Full	18.311413			
Reduced	37.281664			

RSquare (U)	0.5088
Observations (or Sum Wgts)	54

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.66177407	0.8162394	10.63	0.0011
ln(dose)	2.8988935	0.7629513	14.44	0.0001

To estimate the amount of penicillin needed to cure at least half of the rabbits (the ED50), use inverse prediction in either of the following ways:

- Choose the crosshair tool from the Tools menu and move along the dose-response curve until it intersects .5 on the response (Y) axis. Read the approximate ED50 dose, –0.9211, as shown in **Figure B**.
- Click the pop-up menu beneath the plot and choose Inverse Prediction. Enter one or more response rates into the Inverse Prediction dialog and click Done.

The example in **Figure D** shows predicted ln(dose) levels for response rates from .95 to .05. From the Inverse Prediction table you can see that a more accurate estimate of the ED50 is –0.92286, or in original dose units, $e^{-0.92286} = .397381$.

Note: Inverse prediction on the Fit Y by X platform is only available if there are two response levels. If you have more than two levels or have covariates use the Fit Model command.

Mantel, 1963, "Chi-Square Tests With One Degree of Freedom: Extensions of the Mantel-Haenszel Procedure," *Journal of the American Statistical Association,* 58:690-7000.

Figure D Inverse Prediction Dialog and Inverse Prediction Table

Inverse Prediction

Probability	1-Alpha
0.95	0.9500
0.75	
0.5	
0.25	
0.05	
• • •	

Click/Enter values for Probability

Done Help

Enter probabilities and alpha levels. The Inverse Prediction table shows inverse predictions with confidence limits for each combination of probabilities and alpha levels.

Inverse Prediction

Probability	Predicted ln(dose)	Lower Limit	Upper Limit	1-Alpha
0.95000000	0.08912878	-0.33613445	1.30716646	0.9500
0.75000000	-0.54527462	-0.83718753	0.04429014	
0.50000000	-0.92286336	-1.26101445	-0.58175102	
0.25000000	-1.30045209	-1.88351446	-1.00911909	
0.05000000	-1.93485549	-3.14484571	-1.51171724	

STACKING MULTIPLE SETS OF COLUMNS

by Ann Lehman
SAS Institute Inc.

Recently someone asked the question, "What is the easiest way to stack multiple sets of columns?"

For example, what is the simplest way to change this arrangement:

5 Cols / 3 Rows	block	A1	A2	B1	B2
1	1	A11	A21	B11	B21
2	2	A12	A22	B12	B22
3	3	A13	A23	B13	B23

into the following form?

block	A	B
1	A11	B11
1	A21	B21
2	A12	B12
2	A22	B22
3	A13	B13
3	A23	B23

This looks like a job for the Stack command in the Cols menu, but Stack only facilitates stacking into a single output column. This example needs two output columns.

However, there is an easy way to take advantage of several Tables menu and Cols menu features in JMP and get the desired result.

You begin by using Cols→Stack to pile all four columns in the original example (A1, A2, A3, and A4) into a single column, which gives the table shown in **Figure A.**

Figure A Table with One Stacked Column

3 Cols / 12 Rows	block	_Stacked_	_ID_
1	1	A11	A1
2	1	A21	A2
3	1	B11	B1
4	1	B21	B2
5	2	A12	A1
6	2	A22	A2
7	2	B12	B1
8	2	B22	B2
9	3	A13	A1
10	3	A23	A2
11	3	B13	B1
12	3	B23	B2

Next, use Analyze→Distribution of Y to create a histogram of the _ID_ variable produced by the Stack command. Highlight histogram bars to define subsets of data. You do this by highlighting the bars for either the A variable or the B variable as in **Figure B.** (Note: you can also use the Group/Summary command to define subsets.)

Now do two things to the stacked data table:

- Use Tables→Subset to extract the highlighted rows from the table.
- Click the stacked table to again make it the active table, and use Rows→Delete to delete the highlighted rows from it.

You now have the two tables shown in **Figure C.**

The last step in this process is to do a simple *join by row number* of the two subsets. To do this, choose Tables→Join and complete the Join dialog. Also, click the Select Columns box on the Join dialog and keep only the columns shown in **Figure D**.

For the sake of notational convenience, change the column namees in the joined to A and B in the respective subset tables as shown.

Figure D Join Subsets for Final Results

3 Cols / 6 Rows	block	A	B
1	1	A11	B11
2	1	A21	B21
3	2	A12	B12
4	2	A22	B22
5	3	A13	B13
6	3	A23	B23

To summarize

The steps needed to stack multiple sets of columns are:

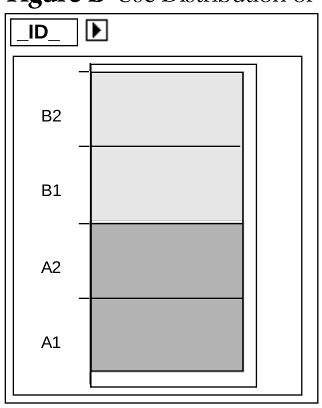
1) Stack all columns into a single column.

2) Create a subset for each set of columns.

3) Join the subsets.

You can use these steps to stack as many sets of columns as required by your data.

Figure B Use Distribution of Y to Identify a Subset



3 Cols / 12 Rows	block	_Stacked_	_ID_
1	1	A11	A1
2	1	A21	A2
3	1	B11	B1
4	1	B21	B2
5	2	A12	A1
6	2	A22	A2
7	2	B12	B1
8	2	B22	B2
9	3	A13	A1
10	3	A23	A2
11	3	B13	B1

Create two subsets:

Highlight histogram bars to identify rows for a subset.

Create the subset of highlighted rows.

Delete the highlighted rows, leaving the original table as the second subset.

Figure C Subsets from Stacked Table

3 Cols / 6 Rows	block	_Stacked_	_ID_
1	1	A11	A1
2	1	A21	A2
3	2	A12	A1
4	2	A22	A2
5	3	A13	A1
6	3	A23	A2

3 Cols / 6 Rows	block	_Stacked_	_ID_
1	1	B11	B1
2	1	B21	B2
3	2	B12	B1
4	2	B22	B2
5	3	B13	B1
6	3	B23	B2

Tips and Techniques

Use The Attributes Table!

There are occasions when many or all the columns in your JMP table need to have the same change made to them. For example, suppose you imported data that had many columns with responses coded as 0 or 1. Survey data with yes/no questions are sometimes coded that way. By default, the JMP import command constructs a numeric data type for a column containing all numbers, and assigns those columns the continuous modeling type.

The table in **Figure A** shows 120 columns called Q1, Q2, . . ., Q120 with continuous numeric 0 or 1 responses. If you wanted a count and a histogram for all Q variables, their modeling type needs to be nominal. Changing each of them through the Column Info dialog would be unreasonably tedious and time consuming.

This is a situation where the Attributes command in the Tables menu comes in handy. The Attributes command creates a new table called an *attributes table* from the active data table, called the *source table*. An attributes table has a row for each column in its source table and a column for each type of column characteristic. You can quickly make global changes to table columns by entering new values into its attributes table:

- Name lists the variable names in the source table. You can type in new column names using up to 31 characters.
- The Type (data type) column has values “Numeric,” “Character,” or “Row State.”
- Measure (modeling type) has values “Continuous,” “Ordinal,” and “Nominal,” which specify how JMP uses the column for data analysis and plotting.
- FW lists the field width of each column.

Figure A Example of Data with Many Columns Characteristics to be Changed

121 Cols / 66 Rows	ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7
1	A	0	1	0	0	0	0	
2	B	0	1	1	0	1	1	
3	C	0	0	0	0	1	0	

Use the Attributes command to change all responses from continuous to nominal and to assign Y Roles.

Figure B Attributes Table Example to Change Modeling Type and Analysis Role

9 Cols / Table Info / 121 Rows	Name	Type	Measure	FW	N Dec	Lock	Source	Validation	Role
1	ID	Character	Nominal	8	•	No	No Formula	None	None
2	Q1	Numeric	Continuous	8	•	No	No Formula	None	None
3	Q2	Numeric	Continuous	8	•	No	No Formula	None	None

9 Cols / Table Info / 121 Rows	Name	Type	Measure	FW	N Dec	Lock	Source	Validation	Role
1	ID	Character	Nominal	8	•	No	No Formula	None	None
2	Q1	Numeric	Nominal	8	•	No	No Formula	None	None
3	Q2	Numeric	Nominal	8	•	No	No Formula	None	None

9 Cols / Table Info / 121 Rows	Name	Type	Measure	FW	N Dec	Lock	Source	Validation	Role
1	ID	Character	Nominal	8	•	No	No Formula	None	None
2	Q1	Numeric	Nominal	8	•	No	No Formula	None	Y
3	Q2	Numeric	Nominal	8	•	No	No Formula	None	Y
4	Q3	Numeric	Nominal	8	•	No	No Formula	None	Y

$ Update Source
Sort Contents…
Add Col Notes
Max Col Notes Width…

121 Cols / 66 Rows	ID	Q1	Q2	Q3	Q4	Q5
1	A	0	1	0	0	0
2	B	0	1	1	0	1
3	C	0	0	0	0	1
4	D	1	1	1	0	1

- N Dec lists the number of decimal places assigned to display numeric columns.
- Lock has values “Yes” or “No,” which tell whether a source table column is locked so it cannot be edited, or unlocked.
- Source has values “No Formula,” “Formula,” or “Instrument” to indicate the source of the input data.
- Validation has values “None,” “Range Check,” or “List Check” to indicate admissible values in a column.
- Role can have values “None,” “X,” “Y,” “Weight,” “Freq,” or “Label.”

Figure B shows the sequence of actions you do to change the modeling types from “Continuous” to “Nominal”, and to assign “Y” to all response columns:

1) Choose Tables→Attributes to see the top table in **Figure B**.

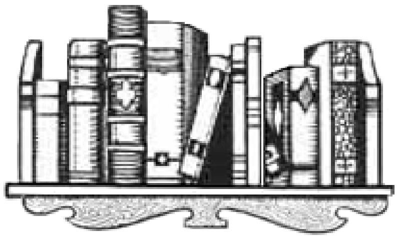
2) Highlight “Nominal” in the Measure column and choose Edit→Copy. Then highlight the Measure column itself and choose Edit→Paste to change all columns to “Nominal,” as shown in the second table of **Figure B**.

3) Enter a “Y” into a Role column cell. Then, select the cell and choose Edit→Copy. Highlight the Role column and all the rows except the first, as shown in the third table of **Figure B**. Choose Edit→Paste to change all response columns to “Y”.

4) Select the Update Source option from the dollar ($) popup menu on the window border to make changes in the source table, as shown at the bottom of **Figure B**.

The attributes table is useful for other commonly encountered global changes such as removing formulas, unlocking columns, and changing data types; and is especially useful to set the same number of decimals to display on multiple columns.

JMP FOR USER PUBLISHING



SAS® User Publishing Program
SAS Institute Inc.

Have you considered writing about JMP? Are you interested in publishing an article or a book? Having difficulty imagining what authoring would be like? SAS Institute's User Publishing Program can assist you!

The SAS User Publishing Program has two branches, one for Books by Users (BBU), the other for *Observations®: The Technical Journal for SAS Software Users.* BBU handles books written by SAS users for SAS users. *Observations* publishes feature articles submitted by SAS users and Institute experts. Both branches are open to authors who want to write about JMP.

The goal of User Publishing is to make the process of writing and publishing as easy for you as possible while ensuring a quality publication. The User Publishing team takes care of details so that you can concentrate on writing. Here is how we can help regarding your JMP publishing needs. We offer:

- complete editorial and publishing support
- technical advice and assistance
- access to an audience of JMP users
- competitive royalties for books
- honoraria for *Observations* articles.

The Author's Packet provides detailed information on how to prepare and submit proposals for BBU and Observations. Proposals include a completed Author's Questionnaire and an outline for a BBU book or a full draft of your *Observations* submission. Your outline or draft will be reviewed by a team of technical experts.

The publishing agreement you sign with the Institute specifies your rights and responsibilities as author and the Institute's rights and responsibilities as publisher. Your acquisitions editor from User Publishing will then assist with development and organization.

During technical review, copyedit and production, your editor will work closely with you to ensure a polished final product. You won't have to focus on grammar, punctuation, usage, spelling or format. Leave the worry to us and soon you'll be a published author!

To request an Author's Packet or additional information regarding BBU or *Observations* send electronic mail to:

`sasbbu@unx.sas.com` or
`sasobs@unx.sas.com`

For more information on either option access SAS Institute's site on the World Wide Web at the following URL:

`http://www.sas.com/`

You may call any of the following BBU representatives at (919) 677-8000:

Irene Down x 6479
`sasimd@unx.sas.com`

Julie Platt x 7447
`sasjmp@unx.sas.com`

Judy Whatley x 7451
`whatley@unx.sas.com`

Betty Angell x 6179
`sasbxa@unx.sas.com`

Hanna Schhoenrock x 6422
`sashxh@unx.sas.com`

Or, send facsimiles to (919) 677-4444
attn: BBU

Mail correspondence to:

User Publishing Program
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513-2414

The User Publishing team will be watching for your JMP submission. Contact us. We'll *jump* for you!

JMP SERVICES UPDATE

Professional Services Division
SAS Institute Inc.

SAS Institute currently offers the following JMP courses:

- Interactive Data Analysis Using JMP Software (2 days)
- Design and Analysis of Experiments Using JMP Software (1 day)
- Engineering Statistics Using JMP Software (2 days)
- Reliability Analysis Using JMP Software (1 day).

SAS Institute is expanding its JMP curriculum to include the following topics (available in November 1996):

- Categorical Data Analysis Using JMP Software
- ANOVA and Regression Methods Using JMP Software
- Multivariate Statistical Methods Using JMP Software
- Advanced Design and Analysis of Experiments Using JMP Software.

You can attend selected JMP training courses at an Institute training center or schedule any of the courses at your site.

In addition to formal classroom training, SAS Institute's Professional Services Division also offers JMP consulting services. You can receive short-term on-site services designed to help you efficiently and effectively use JMP software to complete your data analysis tasks. The Institute can provide you with a full range of expert JMP software services and professional assistance tailored to your organization's immediate computing needs.

For more information on JMP training and consulting services, contact Deborah Upchurch at SAS Institute's headquarters in Cary, NC at

919-677-8000, x 7312.

VARIABILITY AND GAGE R&R ANALYSIS

by Liz Edwards and Dave Tilley
SAS Institute Inc.

JMP 3.1.5 introduces a new feature called Variability Analysis. This kind of analysis is useful for anyone involved in manufacturing, quality control, or any analytical process where frequent measurements are made throughout the process. A manufacturing or analysis process can only produce good quality results if all people and equipment are making consistent measurements.

Human error or damaged equipment can cause poor measurements that contribute to overall product variation. The Variability Analysis platform provides graphical and statistical tools, such as Gage R&R analysis (Gage Repeatability and Reproducibility Analysis), which can help you identify trouble spots in your measurement system.

A *measurement system* includes the actual instruments (sometimes called *gages*) used to make a measurement, the people making measurements (the operators), and the parts (products) being measured. To evaluate a measurement system, a standard set of parts are measured several times by each operator with the different instruments. Once you have collected this data, you can look for variations between instruments, between parts, or between any combination of instruments and parts. In recent years, more quality analysts have recognized the importance of assessing gage capability. There is greater awareness that the tools used to obtain quality measurements are themselves subject to variation. Overall product variation can be broken down into two components: variation due to the product and variation due to the gage. That is:

$$\sigma^2_{total} = \sigma^2_{product} + \sigma^2_{measurement}$$

Gage R&R looks at measurement variation and attempts to quantify the ability of the measurement system to perform adequately.

Repeatability is the ability of the gage to give consistent results when the same technician or operator takes several measurements of the same unit. *Reproducibility* is the ability of the gage to give consistent results when different operators use it to measure the same unit.

Note: The Variability Analysis platform in JMP handles only random effects models with complete and balanced data. That is, there are an equal number of measurements taken at each combination of factor levels, and there are no missing values.

Graphical Methods

To improve a process, you must identify the sources of measurement error and examine the components of variation. **Figure A** shows an example of the arrangement of data for a 2-factor crossed model. The table contains measurements taken by 5 operators on 10 parts. The table, called 2 Factors Crossed (Mac) or 2crossed.jmp (Windows) is in the JMP sample data.

Figure A Variability Analysis Data

3 Cols / 90 Rows	Measurement	Operator	part#
30	0.8	Cindy	10
31	0.65	George	1
32	0.6	George	1
33	0.6	George	1
34	1	George	2
35	1.1	George	2

To produce a Variability chart, you choose Graph→Control Charts, which displays the Control Charts dialog. You then click the Variability Charts button to see the Variability Charts dialog as shown at the lower left of **Figure B**.

The default analysis is a 2-factor crossed model, but other models are available in the Model Type popup menu on the Variability Charts dialog. Enter your response measurements and factors for Operator and Part, and click Chart to see a Variability chart like the one shown in **Figure C**.

Figure B The Variability Charts Dialogs

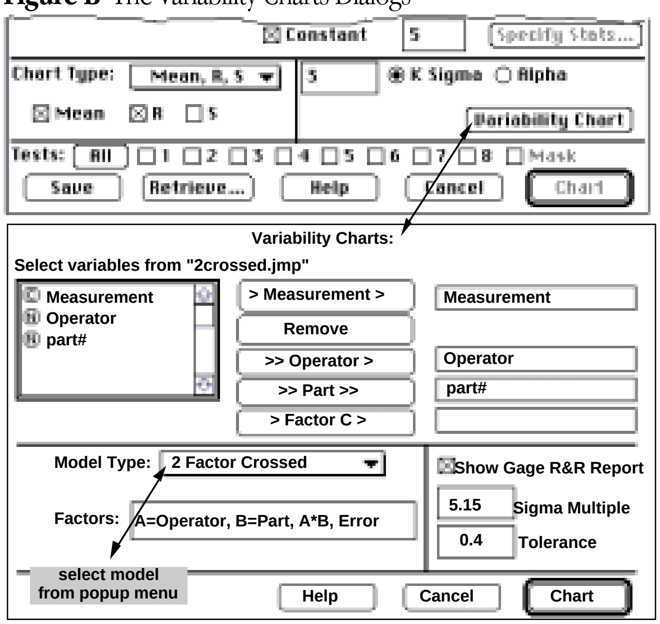
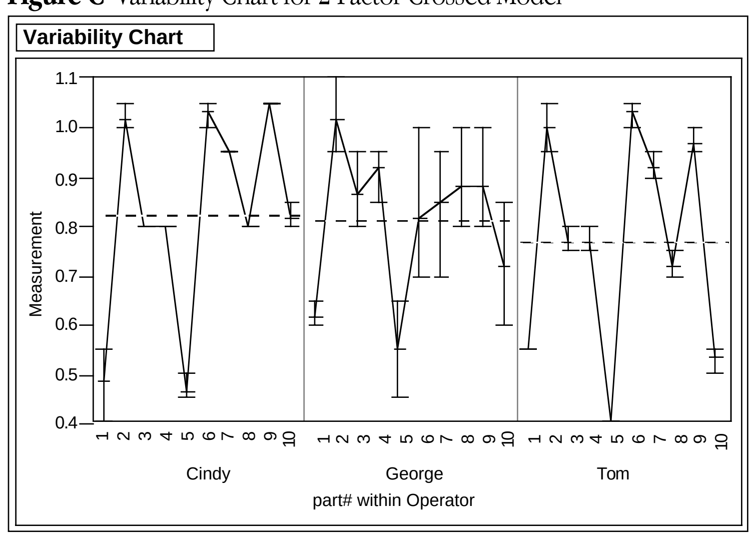


Figure C Variability Chart for 2-Factor Crossed Model



Variability charts offer you a quick visual assessment of several components of variation all at once.

- The error-bar notation at each point shows the range of measurements made by each operator on each part, called “repeatability” in Gage R&R terms. Small ranges for each operator demonstrate the operator’s ability to obtain consistent measurements with the gage.

- You also compare the plotted average of each operator’s measurement on the same part, or “reproducibility”. Close measurements taken by different operators on the same part are a sign of small operator variability. However, you expect the part-to-part measurements taken by each operator to vary substantially. This indicates the gage’s ability to differentiate among parts.

Analysis of Variance and Variance Components

The random effects model estimates variances of effects or variance components.

The Variability Analysis platform also displays the Analysis of Variance and Variance Component Estimates tables for the model type and factors selected (**Figure D**). The “% of Total” column in the Variance Components Estimates table is of particular interest. It shows how the total variation is partitioned among the effects in the model. In this example, most of the variability is attributable to part-to-part variability.

Figure D Analysis of Variance and Variance Components Tables

Analysis of Variance

Source	DF	SS	Mean Square	F Ratio	Prob>F
Operator	2	0.05489	0.02744	1.3150	0.29306
part#	9	2.63358	0.29262	14.0209	0.00000
Operator*part#	18	0.37567	0.02087	5.0425	0.00000
Error	60	0.24833	0.00414		

Variance Component Estimates

Component	Var Comp Est	% of Total	Cum. Total	Cum. %	Sqrt(VC)
Operator	0.00021914	0.55	0.0002	0.5	0.0148
part#	0.03019444	75.24	0.0304	75.8	0.1738
Operator*part#	0.00557716	13.90	0.0360	89.7	0.0747
Error	0.00413889	10.31	0.0401	100.0	0.0643

The Gage R&R Report: Variance Components Method

For a 2-factor crossed design, the Variability Analysis platform displays a Gage R&R Report. JMP uses a statistically more efficient version of the traditional approach to Gage R&R. The calculations are done using estimates of the variance components instead of variance estimates based on ranges; the method of moments approach is used to solve for the variance components. Computations use the sigma multiple (υ) and the tolerance factor you specify in the Variability Chart dialog.

Figure E is the Gage R&R Report generated for this example of a 2-factor crossed model. The Measurement Unit Analysis lists these quantities:

EV Repeatability - equipment variation, computed
$\upsilon \sqrt{\sigma^2_{EV}}$

AV Reproducibility - operator -to-operator variation, computed
$\upsilon \sqrt{\sigma^2_{AV}}$

IV Interaction - interaction of operator and part, computed
$\upsilon \sqrt{\sigma^2_{IV}}$

RR Gage R&R - measure of repeatability and reproducibility, computed
$\sqrt{EV^2 + AV^2 + IV^2}$

PV Part variation, computed
$\upsilon \sqrt{\sigma^2_{PV}}$

TV Total Variation, computed
$\sqrt{RR^2 + PV^2}$

In the Gage R&R report, % Tolerance lists the Measurement Unit Analysis for EV, AV, IV, RR, and PV divided by the tolerance you specified in the Variability dialog.

Discrimination Ratio

The variability of the quality measurement is composed of variance due to the product itself and the variance of the measurement error: You want the measurement process be able to detect product variation. For each factor, the discrimination ratio quantifies the usefulness of the response measurement for a product. (Wheeler and Lyday, 1989).

For a given factor, X, the discrimination ratio compares the product measurement, Y, and variance of the response measurement, V_X, with the measurement error variance, E. The Discrimination Ratio, D, is computed

$$D = \sqrt{\left(\frac{2V_X}{E}\right) - 1}$$

Computational note:
The mean square for error E is taken from the Variance Component Estimates table.

For a factor X with k levels, V_X, is computed

$$\text{var}(\text{mean } Y_k) + ((N_k - 1)/N_k) \bullet E$$

As a rule of thumb:

- A Discrimination Ratio less than 4 indicates the measurement error is so great that it masks the product variation. This means it would be best to work on improving the measurement process

Figure E Gage R&R and Discrimination Reports

Gage R&R

Measurement Unit Analysis		% Tolerance
Repeatability (EV)	0.3313211	82.830
Reproducibility (AV)	0.0762367	19.059
Operator * Part Variation (IV)	0.3846040	96.151
Gage R&R (RR)	0.5133283	128.332
Part Variation (PV)	0.8948923	223.723
Total Variation (TV)	1.0316676	
Sigma Multiple	5.1500000	
Tolerance	0.4000000	

Gage R&R Report assumes that column 'Operator' represents Operator, and column 'part#' represents Part.

Discrimination Ratio

Source	Discrim Ratio
Operator	1.1727709
part#	4.0606574

Guidelines (Barrentine, 1991) for acceptable % RR are:
< 10% excellent
11% - 20% adequate
21% - 30% marginally acceptable
> 30% unacceptable.

because identifying product variation will not even be possible until that is accomplished.

- A Discrimination Ratio greater than 4 implies that it is possible to detect unacceptable variation. (Wheeler and Lyday 1989).

Generalization

The value of variability and Gage R&R analysis is not confined to a particular manufacturing process or industry. For example, the 'operator' factor could be thought of generically as any condition affecting measurements such as a set-up procedure, an environmental condition such as temperature, or date. Likewise the 'part' factor could be any standard factor in a process, such as work shift or weekday.

Variability Analysis in JMP accommodates 1 to 3 factors with interactions, which can be crossed or nested, as well as crossed and then nested.

The flexibility, scope, and ease of use of the variability platform makes it a valuable addition to the statistical and graphical techniques for total quality improvement that are available in version 3.1 of JMP.

ASQC Automotive Industry Action Group (1990), *Measurement Systems Analysis Reference Manual*, ASQC Automotive Industry Action Group.

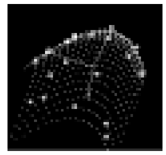
Barrentine (1991), Concepts for R&R Studies, Milwaukee, WI: ASQC Quality Press.

LaBarr, Julie (1994), "SAS Tools for Assessing Gage Repeatability and Reproducibility," *Proceedings of the Nineteenth Annual SAS Users Group International,* Cary, NC, SAS Institute Inc.

SAS Institute Inc., *SAS/QC Software: Usage and Reference, Version 6, First Edition*, Volume 2, Cary, NC: SAS Institute Inc., 1995. 1697 pp.

Wheeler, D.J. and Lyday, R.W. (1989), *Evaluating the Measurement Process., Second Edition*, Knoxville, TN, SPC Press Inc.

VISUALIZING A 3-D FITTED SURFACE



by Michael Hecht
SAS Institute Inc.

JMP's Fit Model platform can produce a wide variety of fits to your data, but some of them are difficult to visualize using JMP's default graphics. In particular, if your data is best fit by a three-dimensional surface, such as a response surface, you must do some extra work to see the fit and the data in the same graph.

I started with the data in **Figure A**: 20 observations of data, variables X, Y, and response variable Z. To identify these original data values, I added the character column named source to the data table, and used a formula to set its value to the constant character string "data".

Figure A Example Data for 3-D Fitting

4 Cols / 20 Rows	X	Y	Z	source
1	-0.1897	-2.9526	0.1911	data
2	-1.5997	3.4883	-0.6412	data
3	-2.8363	-0.0025	0.2998	data
4	2.2537	-1.7853	0.2758	data
5	2.0578	-2.4476	-0.0410	data
6	-1.9189	3.9780	-0.9357	data
7	-1.8569	-0.8812	0.8942	data
8	0.8862	3.5126	-0.4579	data
9	2.5182	1.4841	0.2708	data
10	2.6863	-0.4814	0.3981	data
11	0.3563	2.4833	0.5793	data
12	3.9885	-1.2509	-0.8591	data
13	2.4416	-2.2807	-0.2091	data
14	3.7092	2.9521	-1.0017	data
15	2.9619	2.6976	-0.7708	data
16	0.1880	-1.9412	0.8210	data
17	-1.9135	-2.5845	-0.0128	data
18	3.3541	-2.0509	-0.7098	data
19	3.0623	1.6892	-0.2997	data
20	1.4776	-0.1488	1.0036	data

I wanted to visualize a surface fit to this data in a spinning plot that would show both the original data values and the fitted surface. To begin, I used the Fit Model platform, and produced a fit of Z from a response surface of X and Y.

To get the fitted surface itself, I knew I would have to calculate its points. So, I used the Save Prediction Formula option found in the Save ($) menu of the Fitting Platform. This created the Pred Formula Z column in the data table as shown in **Figure B**, and attached the formula that describes the fit.

Figure B Surface Fit with Predicted Values

$$1.23727277 + -0.0122591 \cdot X + 0.00469837 \cdot Y + X \cdot X \cdot -0.1136486 + Y \cdot X \cdot 0.0083235 + Y \cdot Y \cdot -0.1197364$$

surface fit prediction formula ↓

5 Cols / 20 Rows	Y	Z	source	Pred Formula Z
6	9780	-0.9357	data	-1.09358
7	8812	0.8942	data	0.78467
8	5126	-0.4579	data	-0.2978
9	4841	0.2708	data	0.285899
10	4814	0.3981	data	0.343460

So far, so good. I decided it would be easier to generate the surface points in a separate table. So, I used COMMAND-click (Mac) or ALT-click (Windows) to select (highlight) the X, Y, source, and Pred Formula Z columns. I omitted Z. Then I used Tables→Subset to produce a new data table that contained only those four columns, and renamed Pred Formula Z to simply Z, so it would match the original table. Also, I changed the character string in the source formula to have the value "fit" instead of "data." The subset table looks like this:

Figure C Subset to Generate Surface Value

4 Cols / 20 Rows	X	Y	source	Z
1	-0.1897	-2.9526	fit	0.1825
2	-1.5997	3.4883	fit	-0.5210
3	-2.8363	-0.0025	fit	0.3570
4	2.2537	-1.7853	fit	0.2089

Next, I added formulas for X and Y so they would count in a simple rectangular fashion. I used nested count formulas, as described in Chapter 5, *Calculator Functions*, of The JMP User's Guide.

The formula for X is

$$\text{count}\left(\text{from -3, to 4, in } \sqrt{n} \text{ steps, } \sqrt{n} \text{ times}\right)$$

and Y's formula is

$$\text{count}\left(\text{from -3, to 4, in } \sqrt{n} \text{ steps, 1 time}\right)$$

I picked the start and stop numbers –3 and 4 by using Analyze→Distribution of Y for both X and Y, and choosing the values on the extreme ends of their histograms.

Now I could generate a surface by adding rows to the table.

➡
Continued

➡ **Continued from page 11**

I wanted the number of rows to be a perfect square, so I chose 22x22 = 484 and added enough rows (464) to reach this amount. The number of rows to add is arbitrary, but determines the density of the surface.

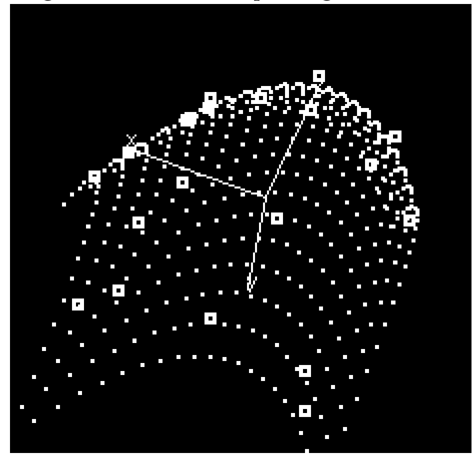
The only remaining task was to glue this table to the original table. I went back to the original table and deleted the Pred Formula Z column, as it was no longer needed. I then used Tables→Concatenate and it produced the final table you see in **Figure D**.

Figure D Combined Data and Fit Table

4 Cols / 504 Rows	X	Y	Z	source
18	3.3541	-2.0509	-0.7090	data
19	3.0623	1.6092	-0.2997	data
20	1.4776	-0.1488	1.0036	data
21	-3.0000	-3.0000	-0.7656	fit
22	-3.0000	-2.6667	-0.5462	fit

At this point, the prize was easily within my grasp! I created a spinning plot of X, Y, and Z. A Distribution of Y histogram of the source variable was useful to select and color or mark the two kinds of points ("data" and "fit"). The result (**Figure E**) was quite impressive (even breathtaking when viewed in color).

Figure E The Prize Spinning Plot



EDITOR
Ann Lehman

CONTRIBUTORS
SAS User Publishing
Duane Hayes
Michael Hecht
Liz Edwards
John Sall
Annette Sanders
Dave Tilley
Deborah Upchurch

TYPOGRAPHY AND DESIGN
Ann Lehman
Jeff Perkinson
Mike Pezzoni

PRINTING
SAS Institute Print Center

JMPer Cable is sent only to JMP users who are registered with SAS Institute. If you know of JMP users who are not registrered, pass them a copy of JMPer Cable and let them see what they are missing!

If you have questions or comments about JMPer Cable, or want to order more copies, write to

JMPer Cable
SAS Institute Inc
SAS Campus Drive
Cary, NC 27513

For More Information on JMP or to order, contact SAS Institute, JMP Sales
phone: 919-677-8000 x 5071
fax: 919-677-4444

You can also browse our web site at
<http://www.sas.com/jmp>

If you don't keep JMPer Cable for reference, please recycle it!

SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513

Address Correction Requested

Bulk Rate
U.S. Postage
PAID
SAS Institute Inc.