

# JMP<sup>®</sup>er Cable



NEWSLETTER FOR JMP<sup>®</sup> USERS



## WHY IS IT CALLED REGRESSION?

Ann Lehman and John Sall  
SAS Institute Inc.

Regression is a method of fitting curves through data points. So why is it called regression?

### History Lesson

Sir Francis Galton, in his 1885 Presidential address before the anthropology section of the British Association for the Advancement of Science (Stigler, 1986), described a study he had made that compared the heights of children with the heights of their parents. He examined the heights of parents and their grown children, perhaps to gain some insight into what degree height is an inherited characteristic. He published his results in a paper, "Regression Towards Mediocrity In Hereditary Stature," Galton, F. (1886).

**Figure A** shows a JMP scatterplot of Galton's original data. The right-hand plot is his attempt to summarize the data and fit a line. He multiplied the women's heights by 1.08 to make them comparable to men's heights and defined the parent's height as the average of the two parents. He defined ranges of parents' heights and calculated the mean child's height for each range. Then he drew a straight line that went through the means as best he could.

He thought he had made a discovery when he found that the heights of the children tended to be more moderate than the heights of their parents.

For example, if parents were very tall the children tended to be tall but shorter than their parents. If parents were very short the children tended to be short but taller than their parents were. This discovery he called "regression to the mean," with the word "regression" meaning *to come back to*.

However, Galton's original regression concept considered the variance of both variables, as does orthogonal regression, which is discussed later. Unfortunately, the word 'regression' later became synonymous with the least squares method, which assumes the X values are fixed.

### A Different Approach

To investigate Galton's situation you can look at the Galton.jmp data table, found in the JMP-IN sample library. Use

**Analyze**→**Fit Y by X** with **child ht** as Y and **parent ht** as X. Select **Fit Line** from the **Fitting** popup menu to see a least squares regression line.

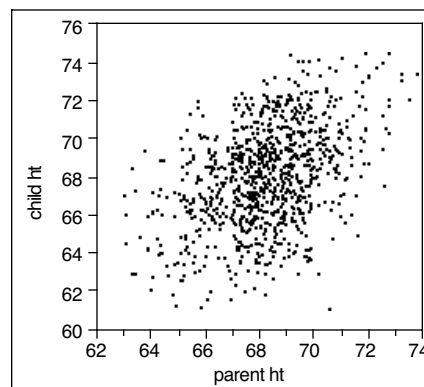
Galton's regression fitted an arbitrary line and then tested to see if the slope of the line was 1. If the line has a slope of 1,

## In This Issue

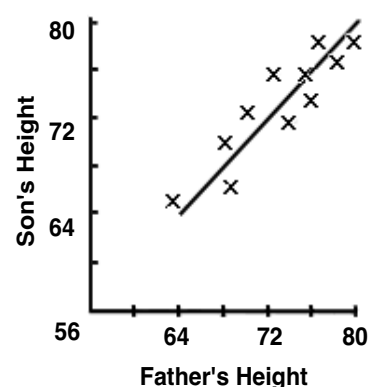
Why is it called Regression? .....	1
To Code Or Not To Code .....	3
JMP Into The Year 2000 .....	6
Calculator Corner .....	7
A Data Table Puzzle .....	7
JMP Data Discovery Conference .....	8
Training Reference Guide .....	8
Compare Normal Quantiles With Standardized Values .....	9
Freq and Weight Variables .....	10
Using JMP to Teach Statistics: A Book Review .....	10
Tips and Techniques .....	11

then the predicted height of the child is the same as that of the parent, except for a generational constant. A slope of less than one indicates *regression* in the sense that the children tended to have more moderate heights (closer to the mean) than the parents. Indeed, the left plot in **Figure B** shows that the least squares regression slope is .61, far below 1, which confirms the regression toward the mean. ➔

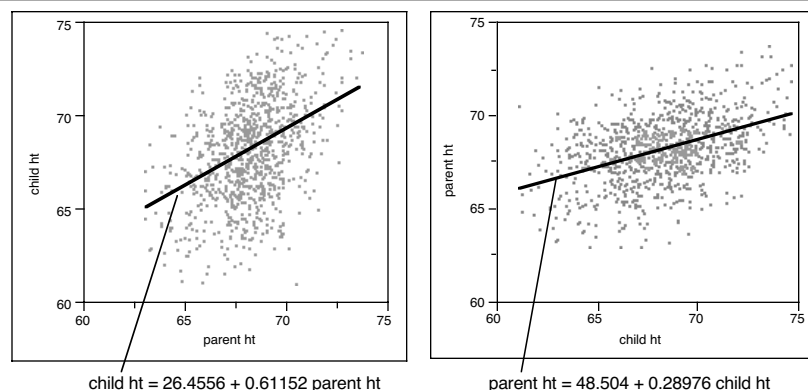
**Figure A** JMP Scatterplot of Galton's Original Data



Galton's Hand-Drawn Scatterplot With Regression Line



**Figure B**  
*Child Height as a Function of  
 Parent (left)*  
 and  
*Parent Height as a Function of  
 Child (right)*



But if the heights of the children were more moderate than the heights of the parents, shouldn't the parents' heights be more extreme than the children's?

To find out, you can reverse the model and try to predict the parents' heights from the children's heights. The analysis on the right in **Figure B** shows the results when **parent ht** is Y and **child ht** is X. If there was symmetry this analysis would give a slope greater than 1 because the previous slope was less than one. Instead it is .29, even less than the first slope.

When you do least squares regression there is no symmetry between the Y and X variables. The slope of Y on X is not the reciprocal of the slope of X on Y; you cannot solve the X by Y fit by taking the Y by X fit and solving for the other variable.

The reason there is no symmetry is that the error is minimized in one direction only—that of the Y variable. So if you switch the roles, you are solving a different problem.

### The Geometry of Linear Regression

An interesting way to visualize regression is to draw a bivariate density ellipse on a scatterplot. The shape and orientation of an ellipse can quickly characterize the relationship of two variables. In fact, Cobb (1998) talks about regression and correlation as *balloon summaries*.

He also uses the density ellipse to graphically illustrate the least squares regression line. On the left in **Figure C** you see a slice of normally distributed points from a scatterplot. For a

given range of X values, a reasonable prediction is the Y value in the vertical slice where the points are the densest—the value under the peak of the normal curve. In fact, this is the least squares prediction. The ellipse in **Figure C** has slices marked at their midpoints. The line through the midpoints of the slices intersects the vertical tangents of the ellipse and is the least squares regression line.

Note that the major axis of the ellipse, which might intuitively seem like it ought to be the regression line, does not cut the midpoints of the slices. For standardized data with X and Y scaled the same, the line along this axis is familiar—it's called the *first principal component*.

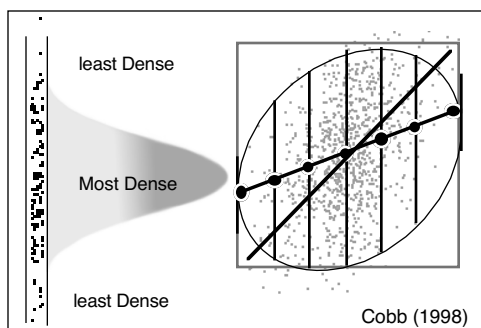
If you reverse the direction of finding midpoints or tangents, you describe what the regression line would be if you reversed the role of the Y and X variables. When you draw the X by Y line fit in the Y by X diagram as shown in **Figure D** it intersects the ellipse at its horizontal tangents.

### Orthogonal Regression

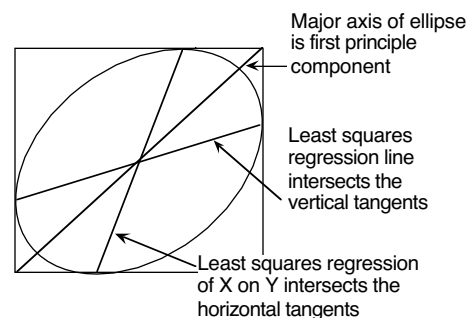
However, there is a way to fit a slope symmetrically, so that the role of both variables is the same. It is called *orthogonal regression*, and uses the ratio of measurement error (error in the X variable) to the response error (error in the Y variable) in equations to estimate intercept and slope parameters (Fuller, 1987). This ratio,  $\sigma^2_X/\sigma^2_Y$ , is zero in the standard least squares regression situation where the variation in X is

(continued on page 5)

**Figure C**  
*Midpoints of  
 Slices Define  
 the Least  
 Squares  
 Regression  
 Line*



**Figure D**  
*Geometry of  
 Linear  
 Regression in  
 Normal Density  
 Ellipse*

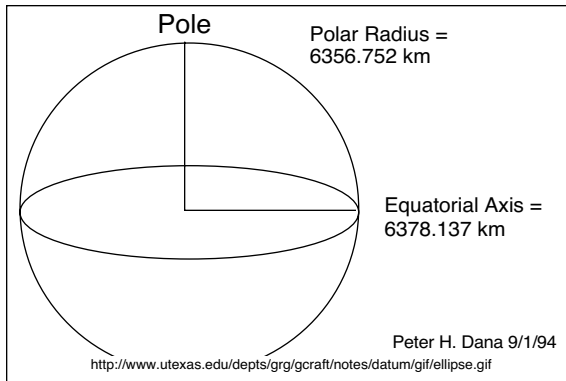


## TO CODE OR NOT TO CODE

Bradley Jones  
SAS Institute Inc.

The earth is not a perfect sphere. As **Figure A** illustrates, the equatorial radius is over 21 kilometers greater than the polar radius. It follows by the Law of Gravity that the same object will weigh more at the North Pole than at the equator.

**Figure A** Earth Radius at Polar and Equator



Armed with this knowledge the intrepid JMP experimental design pedagogy team set out to perform a daring experiment, visiting both the North Pole and Ecuador to measure the force of gravity.

### Experimental Procedure

To measure the force we used a Mettler metric scale precise to 0.001 newton with a digital read-out of 6 digits. We performed two separate weighings of each standard at each location.

We measured the acceleration due to gravity in each location by dropping standard 1 kg and 2 kg masses inside an evacuated 10 meter tube. We timed the fall and calculated the acceleration by solving for  $a$  in the formula:

$$d = \frac{1}{2}at^2 \implies a = \frac{\sqrt{2d}}{t}$$

In each location we repeated this procedure 25 times. The reported acceleration is the average of the 25 measurements and is precise to 0.00001 m/s<sup>2</sup>.

### Data Analysis

Analysis of the data from this study lends insight into the trade-off of using engineering units versus coded units in the regression analysis of experiments.

The data table on the left in **Figure B** shows the data using engineering units. The mass measurements are in kilograms, the acceleration in m/s<sup>2</sup> and the force in newtons. The right-hand table in **Figure B** shows the same data in coded units

**Figure B** Data tables with Factors and Response:

#### engineering units

mass	acceleration	force
1	9.79828	9.798185
1	9.79828	9.798272
2	9.79828	19.5966
2	9.79828	19.59644
1	9.86431	9.864204
1	9.86431	9.864276
2	9.86431	19.72872
2	9.86431	19.72872

#### coded units

mass	acceleration	force
-1	-1	9.798185
-1	-1	9.798272
1	-1	19.5966
1	-1	19.59644
-1	1	9.864204
-1	1	9.864276
1	1	19.72872
1	1	19.72872

units for the two factors, mass and acceleration. The force column remains the same.

The next step is to use **Analyze**→**Fit Model** for each table and run a full factorial least squares analysis. **Figure C** displays the parameter estimates from the least square regression on the two data tables.

**Figure C** Parameter Estimates Tables for Factorial Analyses:

#### engineering units

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.0258327	0.023137	1.12	0.3267
mass	-0.023116	0.014633	-1.58	0.1893
acceleration	-0.002643	0.002353	-1.12	0.3243
mass*accelera	1.0023602	0.001488	673.45	<.0001

#### coded units

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	14.746926	0.000025	600197	<.0001
mass	4.9156915	0.000025	200068	<.0001
acceleration	0.0495521	0.000025	2016.8	<.0001
mass*accelera	0.0165465	0.000025	673.44	<.0001

For the engineering units, only the interaction term (mass\*acceleration) is significant. The coefficient of this term is 1.0023602 with a standard error of 0.001488. The t statistic for testing the hypothesis that the true coefficient is 1 is

$$0.0023602/0.001488 = 1.586$$

with a p value 0.1791. We cannot reject the null hypothesis that the coefficient is 1. This demonstrates that Newton's Second Law of Motion,  $F = ma$ , holds (within our experimental error) both at the North Pole and at the equator.

Because the Intercept, mass, and acceleration terms are not statistically significant, we can also fit a model that contains only the interaction term between mass and acceleration. The results are in **Figure D**. The experimental design, in engineering units, is not orthogonal. In this case fitted coefficients remaining in the model change as the result of excluding terms from the model. Note how the interaction coefficient has changed. It is now 0.9999999 (one for all practical purposes)! This result

**Figure D** Estimates For Uncoded Model Interaction

Parameter Estimates				
Term	Zeroed	Estimate	Std Error	t Ratio
Intercept		0	0	.
mass*accelera		0.9999999	0.000002	491018
				Prob> t
				<.0001

gives even more impressive support for Newton's second law.

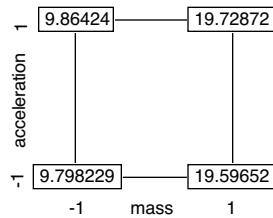
The parameter estimates for the coded analysis (bottom table in **Figure C**) show all the terms, including the Intercept, to be highly significant. How do we interpret this?

Coded units are good for seeing the relative size of factor *effects*. The effect of a factor is the change in the response due to the change in the factor going from its minimum to maximum experimental value.

In the analysis of an orthogonal design using coded units, the regression coefficients are half the factor effect. The coefficients (and factor effects) are directly comparable. Note that the standard error of all the coefficients (0.000025) is the same.

In our experiment the mass coefficient, 4.91 (effect = 9.82), is much larger than the acceleration coefficient, 0.049 (effect = 0.098). The plot shown here graphically illustrates this. (Also see the coded data in **Figure B**.)

The acceleration effect is the average difference between the top and bottom values. The mass effect is the average difference between the right and left values.



The effect of the interaction is still smaller. This is evident by looking at **Figure E**. The 3-D surface plot looks nearly planar. If we could make a greater change in the acceleration, we could observe the twist in the response surface due to the interaction term. (We could achieve this by performing our experiment on the moon or one of the Martian satellites).

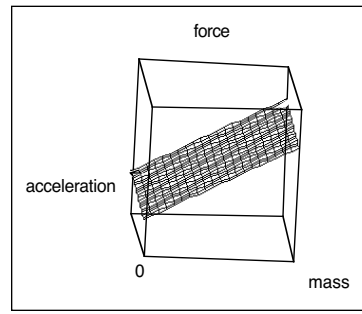
### Trade Off Summary

Engineering Units Pros:

- Coefficients have physical meaning.
- Possible model simplification.

Engineering Units Cons:

- Coefficients are not directly comparable.
- Interaction coefficients are correlated even if the design is orthogonal.
- Forward and backward stepwise regression sometimes terminates with different models.
- Adding terms or removing terms changes the values of coefficients in the model.



**Figure E**

Mesh Plot Illustrates Change of Force For Two Levels of Mass and Acceleration

Coded Units Pros:

- Coefficients are directly comparable.
- All coefficients are uncorrelated if the design is orthogonal.
- Forward and backward stepwise regression always terminates with the same model.
- Adding terms or removing terms does not affect the values of coefficients in the model.

Coded Units Cons:

- Sometimes coding can obscure direct physical interpretation of the data.
- The model can have more terms than absolutely necessary.

In general, coded units are preferable. It is rare that a physical simplification like the one in this article actually happens. Coded units are especially useful in the analysis of screening experiments. In these studies the ability to compare the coefficients directly supports the screening goal. You want to keep the big effects and screen out the small ones. Using coded factors makes effect size comparisons simple.

### Appendix: The Relationship of the Coded and Uncoded Units

What is the functional relationship between the two sets of parameters?

Though the parameter estimates shown in **Figure C** appear quite different, they yield exactly the same predicted values. Regression predictions are invariant to changes in scale in the predictors. This is an important property. It would be disconcerting if we got different predictions by measuring mass and acceleration in English units.

We can use this fact to develop a transformation matrix to change the engineering coefficients to the coded coefficients. Equation 1) shows the standard linear model.

$$1) \quad y = \mathbf{X}\mathbf{b} + \mathbf{e}$$

Equation 2) is the least squares solution for the parameters.

$$2) \quad \mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

In 2) the vector of coefficients is  $\mathbf{b}$  and the superscript,  $\mathbf{t}$ , is the matrix transpose.

Let  $\mathbf{E}$  be the design matrix using engineering units and  $\mathbf{C}$  be the design matrix in coded units. So,

$$3) \quad \mathbf{C}\mathbf{b}_c = \text{predicted force} = \mathbf{E}\mathbf{b}_e$$

We can use 2) and 3) to develop our transformation matrix,  $\mathbf{T}$ , to change the engineering coefficients to the coded coefficients.

$$4) \quad \mathbf{T} = (\mathbf{C}^t\mathbf{C})^{-1}\mathbf{C}^t\mathbf{E}$$

$$5) \quad \mathbf{b}_c = \mathbf{T}\mathbf{b}_e \quad \text{and} \quad \mathbf{b}_e = \mathbf{T}^{-1}\mathbf{b}_c$$

Note: See “The Calculator Corner” in this issue of JMPer Cable for formulas that convert uncoded values to coded value.



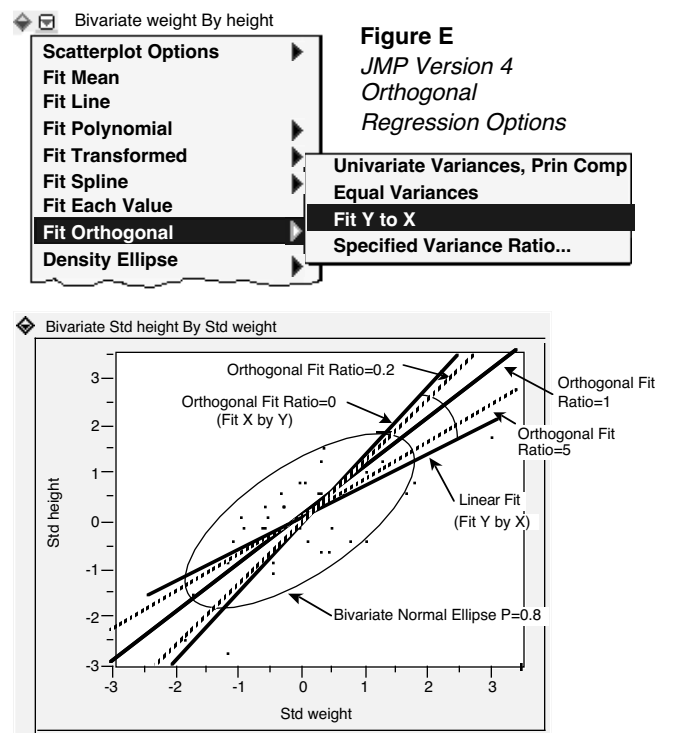
(continued from page 2)

ignored or assumed be zero, and becomes infinitely large when the variation of  $Y$  approaches zero. An interesting advantage to this approach is that the computations give you predicted values for both  $Y$  and  $X$ .

**Orthogonal Regression** is a new fitting option on the Fit  $Y$  by  $X$  platform and will be available in JMP Version 4 with the following options (see **Figure E**) to specify a variance ratio:

- **Univariate Variances, Prin Comp**  
uses the univariate variance estimates computed from the samples of  $X$  and  $Y$ .
- **Equal Variances**  
uses 1 as the variance ratio. If the variables are already standardized the fitted line represents the first principle component, as illustrated previously in **Figure D**.
- **Fit  $X$  to  $Y$**   
uses a very large variance ratio, which indicates that  $Y$  has effectively no variance (see **Figure D**).
- **Specified Variance Ratio**  
lets you enter any ratio you want, giving you the ability to make use of known information about the measurement error and response error.

The scatterplot in **Figure E** shows standardized height and weight values with various line fits that illustrate the behavior of the orthogonal line selections. The standard linear regression occurs when the variance of the  $X$  variable is considered to



be zero. The Fit  $X$  by  $Y$  is the opposite extreme, when the variation of the  $Y$  variable is ignored.

All other lines fall between these two extremes and shift as the variance ratio changes. As the variance ratio increases, the variation in the  $Y$  response dominates and the slope of the fitted line shifts closer to the  $Y$  by  $X$  fit. Likewise, when you decrease the ratio, the slope of the line shifts closer to the  $X$  by  $Y$  fit.

A biographical note: Galton was the cousin of Darwin and mentor of Karl Pearson. This British statistician was also an explorer, and anthropologist, and perfected an early technique for fingerprinting. (The first legal use of fingerprints was in the conviction of a billiard ball thief in 1902.)

#### References

- Cobb, G.W. (1998), *Introduction to Design and Analysis of Experiments*, Springer-Verlag: New York.
- Galton, F. (1886), “Regression Towards Mediocrity in Hereditary Stature,” *Journal of the Anthropological Institute*, 246-263.
- Fuller, W. A. (18987), *Measurement Error Models*, John Wiley & Sons, New York
- Stigler, S.M. (1986), *The History of Statistics*, Cambridge: Belknap Press of Harvard Press.



# JMP INTO THE YEAR 2000

by Michael Hecht  
SAS Institute Inc.

JMP software, beginning with Release 3.0 is year 2000 compliant under both Macintosh Operating System and Windows. JMP stores dates with four-digit values; you can manage and manipulate data involving dates without worrying about erroneous results. If you enter a two-digit date, JMP uses the algorithms supplied by the Mac OS date/time utilities to assign the century.

**Note:** If you are using JMP under Windows, date and time values are interpreted in the exact same manner.

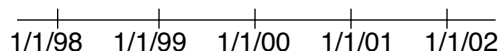
## Using Date Formats Within JMP

When you assign a date format to a numeric column in JMP it is stored internally as a signed, 64-bit integer representing the number of seconds since midnight on January 1, 1904. This enables date columns to be sorted and compared.

As a user, you decide how to display the date. A *long* date format displays as "Wednesday, January 01, 1997". A *short* date format displays as 1/1/97 (dd/mm/yy). Additional formats are available for both date and time values.

## Entering 4-Digit Years

If you enter or import a date value with four digits in the year, JMP stores the date correctly and displays either two or four digits of the year depending upon the format you select. For example, if you use a short date format, a sorted list of dates for the 20th and 21st centuries might look like this on an axis of a graph:



1/1/98   1/1/99   1/1/00   1/1/01   1/1/02

Because of their internal representation, the dates appear in their sorted order with the year 1999 appearing before the year 2000.

## Working With 2-Digit Years

So what happens if you enter a two-digit year? How does JMP decide if the date is in the 20th or 21st century?

JMP relies on the Mac OS date and time text conversion utilities to assign the correct century.

By relying on the Mac OS utilities, JMP does not need to worry about year 2000 compatibility. As Apple Computer Inc. states in their Apple Directions newsletter of September 1996, "All Mac OS date and time utilities have correctly

handled the year 2000 since the introduction of the Macintosh computer." The current Mac OS date and time utilities use a 64-bit signed value to store dates from 30,081 B.C. to 29,940 A.D.

## Assigning A Century For A Two-Digit Year

The Mac OS chooses a century for a two-digit year based on the current year and the value of the two-digit year (which we will refer to as the input year). Mac OS applies specific rules to the input year when the current year falls within a decade of the century.

- If the current year is between 1990 and 1999 inclusive and the input year is 10 or less, the year is assigned to the 21st century. Mac OS assumes that you are referring to a date in the year 2010, for example, rather than in 1910.
- Conversely, if the input year is greater than 10, the year is assigned to the 20th century. For example, if you enter 70, as the two-digit year, Mac OS assumes you are referring to a date in the year 1970 rather than in 2070.

Let's assume the current year is 1998:

The date 1/1/98 is interpreted as 1/1/1998.

The date 1/1/10 is interpreted as 1/1/2010.

The date 1/1/11 is interpreted as 1/1/1911.

Once the 21st century rolls in, the Mac OS applies a similar rule for interpreting two-digit years.

- If the current year is between 2000 and 2010 inclusive and the input year is greater than or equal to 90, the year is assigned to the 20th century.
- If the input year is less than 90, the year is assigned to the 21st century.

Let's assume the current year is 2000.

The date 1/1/98 is interpreted as 1/1/1998.

The date 1/1/10 is interpreted as 1/1/2010.

The date 1/1/11 is interpreted as 1/1/2011.

What happens when the current year is 2011? All two-digit input years will be assigned to the 21st century. This rule stays in effect until the current year falls within a decade of the next century. Therefore in the year 2090, the Mac OS begins applying specific rules, as stated above, for determining the correct century.

## For More Information

For information on the Mac OS utilities for date and time manipulation, see Chapter 4, "Date, Time, and Measurement Utilities," of Inside Macintosh: Operating System Utilities. Also refer to Chapter 5, "Text Utilities", of Inside Macintosh: Text.

next page ➔

To access the September 1996 Apple Directions newsletter, link to :

<http://devworld.apple.com/mkt/informed/appledirections/sep96/year2000.html>

## References

Apple Computer, Inc. Inside Macintosh: Operating System Utilities. Reading, Mass.: Addison-Wesley Publishing Company, 1994.

Apple Computer, Inc. Inside Macintosh: Text. Reading, Mass. Addison-Wesley Publishing Company, 1993.



## Calculator Corner

by Mark Bailey  
SAS Institute Inc.

### CHANGE UNCODED DESIGN FACTOR VALUES TO CODED VALUES

Earlier in this issue the article "TO CODE OR NOT TO CODE" showed a JMP data table of uncoded (engineering) factor values in a design of experiments example,, and a second table of coded values.

A situation could arise where you obtain a table of engineering values and want a simple way to convert them to their corresponding coded values. A way to do this is to add a column to the data table for each factor that contains values in engineering units. Then compute the coded values with the calculator.

The formula in **Figure A** shows how to generate coded values. You use this formula for each factor in the model. Note that the use of the quantile function in the transformation assumes that the minimum value in the variable is equal to -1 and the maximum is 1.

**Figure A**

Formula to Convert  
Uncoded (Engineering)  
Factor Values  
To Coded Values

$high \leftarrow \text{quantile}_1 \text{ acceleration}$

$low \leftarrow \text{quantile}_0 \text{ acceleration}$

$midrange \leftarrow \frac{high + low}{2}$

$halfrange \leftarrow \frac{high - low}{2}$

results  $\frac{acceleration - midrange}{halfrange}$

5 Cols				
	mass	acceleration	force	cancel
1	1	9.79828	9.798185	-1
2	1	9.79828	9.798272	-1
3	2	9.79828	19.596599	1
4	2	9.79828	19.596438	1
5	1	9.86431	9.864204	-1
6	1	9.86431	9.864276	-1
7	2	9.86431	19.728715	1
8	2	9.86431	19.728717	1

Note: For two-level designs, you can analyze the uncoded values as nominal variables and get the same result as with coded values

## A DATA TABLE PUZZLE

The puzzle is, "How do you get from the table on the left in **Figure A** to the table on the right?"

3 Cols		
	B	A
1	B1	A1
2	B1	A1
3	B1	A1
4	B2	A1
5	B2	A1
6	B2	A1
7	B1	A2
8	B1	A2
9	B2	A2
10	B2	A2
11	B1	A3
12	B2	A3

**Figure A**  
How Do You Get From  
Here to Here

4 Cols			
	B	A1	A2
1	B1	1	7
2	B1	2	8
3	B1	3	9
4	B2	4	10
5	B2	5	10
6	B2	6	10

Your first try might be to use **Tables→Split Columns** with **Y** as the Split variable and **A** as the Column ID. That operation selects the first value it finds for each level of **A**, and creates the first row you see in the table shown here. The second row is the second value found for each level of **A**. There is no third

4 Cols			
	B	A1	A2
1	B1	1	7
2	B1	2	8
3	B1	3	9
4	B2	4	10
5	B2	5	10
6	B2	6	10

value for each level of **A** so you see a missing value for **A3** in the third row. Try as you will, there is no allocation of variables in the Split Columns dialog that gives the table you're aiming for in **Figure A**. The table you want has to take the values of the **B** variable into consideration but this approach doesn't tell the split operation anything about **B**.

The trick is in knowing that the Split Columns operation honors a summary table with By-Mode turned on. You begin by choosing the **Group/Summary** command from the **Tables** menu, select **B** as the Grouping variable and click **OK**. To see the table shown here:

- Select **By-Mode On** from the popup menu in the upper left corner of the *summary* table.

2 Cols	
	N Rows
1	B1
2	B2

- Highlight all rows in the table. The highlighted rows are linked to the rows in the original data table (the *source* table) and define the rows corresponding to each level of **B** as subsets.

Then, with the summary table active, use the **Split Columns** command as described above. This time, the split operates on each defined subset independently and concatenates the result. In fact, you could produce the same result—first create a real subset for each level of **B**, then split **Y** in each subset with **A** as the column ID, and use the **Concatenate** command to combine the results.







"You are today's pioneers," said Colleen Jenkins, Director of Statistical Instruments for JMP Statistical Discovery Software, as she addressed the attendees at the opening of the JMP Data Discovery Conference. Held October 28-31, 1997, at SAS Institute's headquarters in Cary, NC, the conference was attended by JMP users from various backgrounds and organizations around the country. During the four days of interactive training they were encouraged to "be like detectives" looking at all the clues and all the angles of their data, by John Sall, Senior Vice President and co-founder of SAS Institute Inc. Sall also leads the JMP software development team. Specifying "Seven Steps to make a JuMP," John stressed the importance of exploring, visualizing, and finding graphs to help tell a story. "The more interaction you have with your data, the more you'll find out about your data," he said.

Bob Stuart, a subject matter expert in the College of Technology at Motorola University, was the Invited Speaker. Bob spoke on *Integrating Statistical Thinking in the Culture*. "JMP makes it quite easy to use the teaching method I like best—visualization," he said. Relating a story of how his grandson learned to swim as analogous to how all learning should be approached, Bob stated that his goal for teaching is to make it as easy and as much fun as possible. He said that JMP software helps him do that.

Aside from providing a forum to learn key statistical methods and analyses using JMP software and sharing experiences with other attendees, highlights of the conference included a dinner at Prestonwood Country Club, a Southern-style breakfast, and a tour of the SAS Institute campus. In addition, the JMP Technology Tour allowed attendees to interact with JMP software developers and Institute staff as well as get a chance to preview Version 4.0 of JMP.

Plan to attend the fall JMP Data Discovery Conference at SAS Institute in Cary from October 6 to October 9, 1998.

Dr. Mark Bailey will be the featured speaker. He is a Statistical Services Specialist at SAS Institute Inc. who received his Ph.D. in chemistry from the University of Rochester. Mark has more than 15 years of experience in research and development of medical diagnostic products, first at Eastman Kodak and later at Abbot Laboratories. In addition, he is a five-year Black Belt for division-wide Six Sigma initiative.

For more information about the upcoming conferences or to

register, call

919-677-8000 X5005

or send a FAX to

919-677-8225



## PROFESSIONAL SERVICES DIVISION TRAINING REFERENCE GUIDE

The professional services division offers the following types of Instructor-based training:

### Public Training

Customer attends a JMP software training course at one of the many Institute training facilities throughout the United States and Canada. Individuals who register and pay for a public training course at least four weeks prior to the course starting date receive a 15% discount. Individuals providing written verification of affiliation with a post-secondary degree-granting institution receive a 50% discount on SAS Institute training courses.

### On-Site Training

Customer requests that SAS Institute provide JMP training courses for their company at their facility.

### Course Modules

- Statistical Data Exploration Using JMP Software
- ANOVA and Regression Methods Using JMP Software
- Categorical Data Analysis Using JMP Software
- Reliability Analysis Using JMP Software
- Statistical Quality Control Using JMP Software
- Design and Analysis of Experiments Using JMP Software
- Multivariate Statistical Methods Using JMP Software

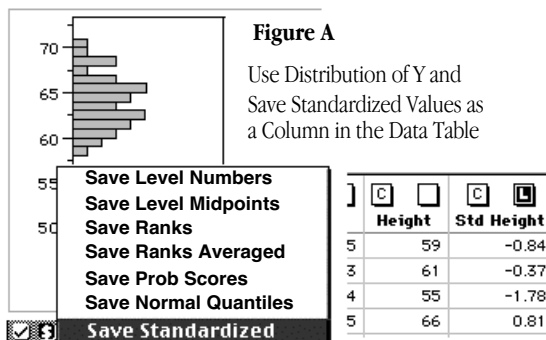




# COMPARE NORMAL QUANTILES WITH STANDARDIZED VALUES

by Nicole Jones  
SAS Institute Inc.

An interesting way to investigate the normality of a distribution is to graphically compare the standardized data values with their corresponding normal quantile values. To do this, first select **Analyze**→**Distribution of Y** for the variable you want. (This example uses the **Height** variable from the Big Class sample data table.) Then choose **Save Standardized** from the dollar (\$) menu on the lower left window border, which computes the standardized **Height** values and saves them as a new column called **Std Height** (**Figure A**).



**Figure A**

Use Distribution of Y and  
Save Standardized Values as  
a Column in the Data Table

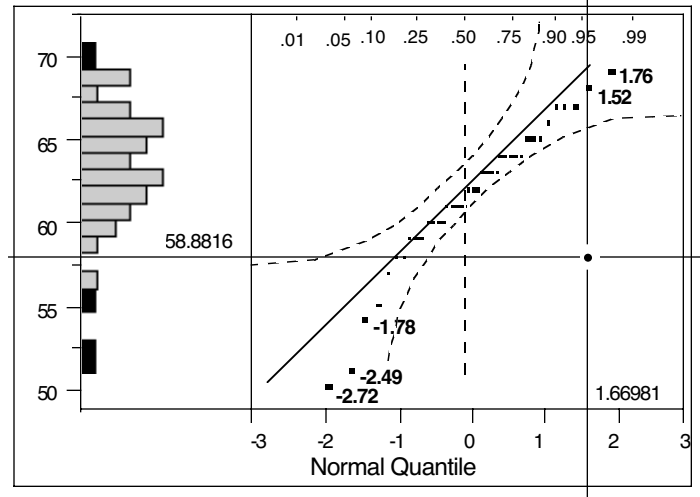
Do the following to see a comparison of standardized values and normal quantiles in the Normal Quantile plot given by the Distribution of Y platform:

- Set the role of **Std Height** to Label with the role assignment popup menu at the top of that column.
- Highlight (CONTROL-click or ALT-click) any points you want to look at in the Normal Quantile plot.
- Select **Rows**→**Label/unlabel** to see the standardized height values on the Normal Quantile plot.
- Select the crosshair tool from the **Tools** menu and align the crosshair vertically with a point to display its normal quantile value as shown in **Figure B**. Move the crosshair from point to point and note the difference between the standardized values and the normal quantile values.

Also, note that you can save the normal quantile values and compute the difference between them and the corresponding standardized values to see the exact difference.

For a theoretical normal distribution the normal quantile value and the standardized value for a data point are the same. So looking at the difference between them can give you a point by point sense of the nonnormality of a distribution.

**Figure B** Label Normal Quantile Points With Exact Standardized Scores and Compare With Crosshair Tool



## Computations

The standardized values can be computed with the calculator as

$$\frac{\text{Height} - \overline{\text{Height}}}{\text{std Height}}$$

The normal quantile values are computed as

$$\Phi \left( \frac{r_i}{(n + 1)} \right)$$

where

$\Phi$  is the normal cumulative distribution function (the NormQuant function in JMP),

$r_i$  is the rank of the  $i$ th observation,

$n$  is the number of nonmissing observations.

You can **Save Normal Quantiles** in the data table and also compute them yourself as a comparison. To compute them, first **Save Ranks Averaged**, (see **Figure A**) then use the calculator formula:

$$\text{normQuant} \left( \frac{\text{RankAvgd height}}{(n + 1)} \right)$$

These normal scores are Van der Waerden approximations to the expected order statistics for the normal distribution.

Note: there are several ways in JMP to examine the normality of a set of values. The simplest way is to use **Distribution of Y** in the **Analyze** menu and look at the shape of a distribution. You can get more exact information with the **Test Dist is Normal** option, which calculates a Shapiro-Wilk W test if  $n \leq 2000$ , or a KSL test if  $n > 2000$ , and reports the probability. If the p-value is less than .05 (or the alpha you are interested in) then you conclude that the distribution is not normal.



## FREQ AND WEIGHT VARIABLES

The **Freq** role identifies a variable whose values are frequencies for each row. Specifically, if  $f$  is the value of the Freq variable for a given row, then that row is used in computations  $f$  times.

Normally, when there is no Freq column, each row contributes the implied value of one (1) to the frequency count. The total number of rows,  $n$ , is the denominator of the mean computation, and is used in determining degrees of freedom for tests of hypotheses.

An analysis that uses a Freq variable reflects an expanded number of rows; means and degrees of freedom reflect this expanded table. You could produce the same analysis (without a Freq column) by first creating a new data table that contains the number of observations expanded by the value of the Freq variable for each row. For example if the value of Freq variable is 5 for the first row, then the first 5 rows in the new expanded table would be copies of the first row of the original table.

You usually think of a Freq value as being an integer greater than 1. However, there are situations where its value has a fractional part, or may be between zero and one. If the value is not an integer, only the Floor of the value is used. If the value of Freq is missing or is less than one, the floor of its value is zero and that observation is not used in the analyses.

Note: A negative frequency value causes an error message. If all frequencies are less than one, the frequencies for computations are all zero and generate an error condition.

### Freq vs. Weight

You assign a column the **Weight** role when you want the analysis to use relative weights for each row. The response in each row is multiplied by its weight variable for all analyses. The sum of the weights is used in statistical computations, however the weight variable does not alter the degrees of freedom.

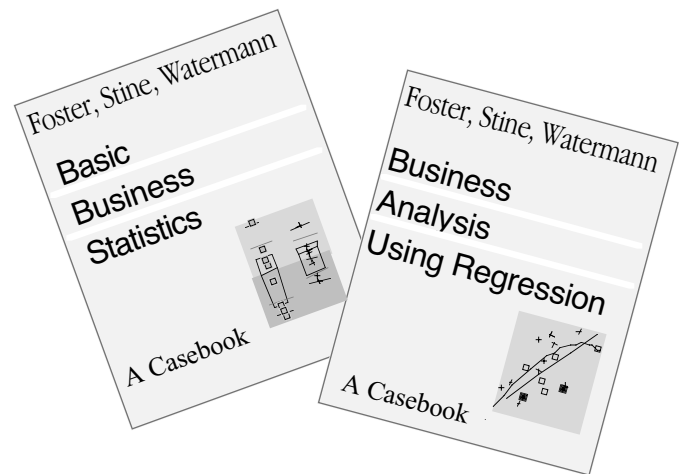
One common use for a Weight variable is to compute weighted product-moment (Pearson) correlation coefficients. A weight variable with values proportional to the reciprocals of the variances is sometimes used when the variance associated with each row is different.

Unlike the Freq variable, the values of a Weight variable can be either integer or noninteger, positive or negative.



## USING JMP TO TEACH STATISTICS

*book review*



Copyright 1998—fresh off the press. Dean P. Foster, Robert A. Stine, and Richard P. Waterman are members of the Department of Statistics, at the Wharton School, University of Pennsylvania. They have written two remarkable books designed as teaching and learning tools for professors of statistics and their students.

The challenge to write a *really good* statistics book has, without question, been met by the two casebooks, *Basic Business Statistics* and *Business Analysis Using Regression*. These books begin with the basics and, using JMP as the analysis software, take the student through simple and multiple regression, analysis of variance and covariance, categorical analysis, and basic time series concepts.

You expect all this in good statistics books. But what jumps up and captivates you is the fun and clever collection of data used for the case studies. For example, a certain brand of chocolate chip cookie claims to have “1,000 chips in every bag.” The students job is to find out if that is true (statistically speaking). Another chip study looks at the thickness of computer chips in a manufacturing process. And of course, there are tables called Forbes, Portfol, Stocks, and Mutfunds. Other interesting table titles are Orings, Flextime, Salary, Utopia, Juice, Headache,...

These books are an organized collection of case studies, wrapped up in large (8 1/2 x 11), easy to read (lots of white space), soft bound (not too heavy) manuals. They are a treasure, an asset in any library.

Dean P. Foster, D.P., Stine, R.A. Stine, and Waterman, R. P. (1998), *Basic Business Statistics : A Casebook*, Springer-Verlag: New York.

Dean P. Foster, D.P., Stine, R.A. Stine, and Waterman, R. P. (1998), *Basic Business Analysis Using Regression: A Casebook*, Springer-Verlag: New York.

# Tips and Techniques

## 5,000 COLUMNS AND COUNTING...

As computers get bigger and faster, and memory costs less and less, JMP is being used more and more to process bigger and bigger data tables. Recently, JMP technical support addressed a request to find a simple way to delete every other column from a data table that had 5000 columns. This means selecting (highlighting) every other column—obviously, clicking 2,500 column headings with the mouse doesn't seem like a reasonable approach.

This problem is easily solved with the Attributes table. When you select **Tables→Attributes**, JMP creates a new table, called an *Attributes* table, from the active data table, called the *Source* table. The Attributes table has a row for each column in its Source table and a column for each type of column characteristic.

An Attributes table is linked to its Source table. You can modify the characteristics of the source table columns by editing values in corresponding Attributes table rows. The table on the left in **Figure A** shows an example table with 100 rows and 5,000 columns (call them X1-X5000). Its Attributes table looks like the table on the right, with 5000 rows, and 9 columns of source table column characteristics.

You can manipulate the Source table in the same ways as any other data table. There is no effect on the Source table until you select the **Update Source** command from the dollar (\$) menu at the lower left of the Attributes table.

This example proceeds with the following steps:

- 1) With the Attributes table active, select **Cols→New Column**, and specify **Row State** as its Data Type and **Formula** as its Data Source.

- 2) Use this formula to identify every other row in the Attributes table (thus identifying every other column in the Source table):

$\text{select}((i \bmod 2) = 1)$

**Figure B** Use a Formula to Select Alternate Rows

Table Info	Source	Validation	Role	Column 10
10 Cols 5000 Rows				
1	No Formula	None	None	■
2	No Formula	None	None	■
3	No Formula	None	None	■
4	No Formula	None	None	■
5	No Formula	None	None	■
6	No Formula	None	None	■
7	No Formula	None	None	■

- 3) To see the table in **Figure B** use the **Copy to Row State** command in the popup menu at the top of the new column. All of the odd numbered rows are selected.
- 4) Choose **Rows→Delete Rows** to delete the selected rows.
- 5) Then select the **Update Source** command in dollar (\$) popup menu at the bottom-left of the table (See **Figure A**). The resulting example table is shown in **Figure C**.

**Figure C** Update Source Table To Remove 2500 Rows

EXAMPLE .JMP					
Table Info	X2	X4	X6	X8	X10
2500 Cols 100 Rows					
1	12	14	16	18	110
2	22	24	26	28	210

You can use the Attributes table to subset columns based on any column criteria. For example, to create a subset with only the numeric columns, create an Attributes table, then:

- generate a histogram (**Analyze→Distribution of Y**) for the variable **Type** and shift-click the “Character” and “Row State” histogram bars to highlight rows in the Attributes table
- delete the highlighted rows and update the Source table.

Using an Attributes table in this fashion gives you a powerful tool for manipulating a very large data table.

EXAMPLE .JMP		
Table Info	X1	X2
5000 Cols 100 Rows		
1	11	12
2	21	22

Attributes of EXAMPL		
Table Info	Name	Type
9 Cols 5000 Rows		
1	X1	Numeric
2	X2	Numeric
3	X3	Numeric

**Figure A**

*Example of Source Table with 5000 columns and Its Associated Attributes Table with 5000 rows*



**EDITOR**

Ann Lehman

**CONTRIBUTORS**

Mark Bailey

Michael Hecht

Bradley Jones

Nicole Jones

Ann Lehman

John Sall

**TYPOGRAPHY AND DESIGN**

Ann Lehman

Mike Pezzoni

**PRINTING**

SAS Institute Print Center

© Copyright 1998 SAS Institute Inc.  
All rights reserved.

JMPer Cable is sent only to JMP users who are registered with SAS Institute. If you know of JMP users who are not registered, 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 a copy, contact SAS Institute, JMP Sales

phone: 919-677-8000 x 5071  
FAX: 919-677-8224

**JMPer Cable is on the Web**

You can now see JMPer Cable at the  
JMP Web site:

[<http://www.jmpdiscovery.com>](http://www.jmpdiscovery.com)

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

SAS, JMPer Cable, and JMP are registered trademarks of SAS Institute Inc. Other brand and product names are registered trademarks or trademarks of their respective companies.

**MARK YOUR CALENDAR FOR THE NEXT JMP DATA DISCOVERY CONFERENCE**

Attend the fall JMP Data Discovery Conference at SAS Institute in Cary from October 6 to October 9, 1998.

Dr. Mark Bailey will be the featured speaker. He is a Statistical Services Specialist at SAS Institute Inc. who received his Ph.D. in chemistry from the University of Rochester. Mark has more than 15 years of experience in research and development of medical diagnostic products. In addition, he is a five-year Black Belt for division-wide Six Sigma initiative.

For more information about the upcoming conferences or to register, call

919-677-8000 X5005 or send a FAX to 919-677-8225



SAS Institute Inc.  
SAS Campus Drive  
Cary, NC 27513

*Address Correction Requested*

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