

Paper 298-2011

Using JMP® Visualizations to Build a Statistical Model

George J. Hurley, The Hershey Company, Hershey, PA

ABSTRACT

JMP has long been used by engineers to build various types of models, including statistical models. However, many statisticians and business analysts have chosen to remain within the SAS/STAT® package to develop their statistical models, possibly due to the historical relationship between statisticians, the pharmaceutical industry, and FDA submissions. However, statisticians are becoming much more common in the business analytics fields, including marketing and pricing. Given that these areas do not have the stringent coding guidelines of FDA oversight, it is natural that statisticians should evolve and begin using some of the more interactive and insightful tools available in conjunction with traditional SAS/STAT. JMP is the perfect tool to add to the portfolio. This is because the visualization elements of JMP and simple interface make developing statistical models highly intuitive with JMP. While JMP is not a replacement for SAS/STAT, it should be considered an essential addition to any statistical software portfolio. This paper will discuss in detail the graphical features in JMP that are particularly useful in developing a linear model.

INTRODUCTION

Statistics is a relatively new profession in comparison to many others. While theoretical statistics dates back to the 17th and 18th centuries (Gauss is credited with developing the method of least squares in 1794), it is only recently, with the development of modern computing and the ability to collect large amounts of data, that the field has been able to blossom as a profession that now is integral to a plethora of fields including agriculture, pharmaceuticals, government, marketing, pricing, and general business strategy. In fact, statistical computing as field didn't exist until the second half of the 20th century, and there are people attending the SAS® Global Forum that can remember the early history of SAS and the release of SAS 76.

Some of the initial uses and popularity of SAS were in the agricultural and pharmaceutical industries. However, SAS has grown to include a host of products that range from business insights tools to advanced data and text mining. In fact, the scope of SAS Software has never been larger. Not only has SAS grown in its breadth of abilities, but it has also grown in its depth. Within this statistical depth are not only included more and more sophisticated statistical procedures, but also simpler, more intuitive ways of doing things. As SAS has evolved, we in the industry should also evolve our thinking to make use of what is available. In particular those in the business fields do not have the limitations placed on analysis tools and coding by the FDA. In consideration of this, an effort should be made to make use of the full compliment of tools available. JMP is a great example of a tool that integrates perfectly with Base SAS and through the use of its visual components, can make the development of statistical models much more streamlined.

In detail, JMP Graph Builder, Distribution Analysis, Model Fitting, Statistical Graphics, Prediction Profiler, and Interaction Profiler will be discussed in detail. While JMP's visualization capabilities extend well beyond just these features, these features represent a sufficient set of the functionality of JMP with respect to visualization, such that the ease of JMP statistical model development will be easily illustrated.

DATA

While JMP and JMP Pro provide excellent visualization and analytic capabilities, it is still, in this author's opinion, not a replacement for the flexibility that is offered in data manipulation by Base SAS... nothing may ever be. With this, it is recommended that a synergism of JMP and SAS be used. This can be accomplished in three ways. First, JSL (JMP Scripting Language) can be used to "SASSUBMIT" a SAS program and then import it and open it in JMP. Secondly, a SAS Program can be run within JMP, but selecting File, then SAS, then New SAS Program.

However, a connection must be created within JMP to SAS to use either of these methods. To connect to SAS, go to File, SAS, Server Connections, and choose "Connect to SAS on this machine," and then click the "Connect" button. You then click the "Close" button on the window. It should be noted that either you need SAS on your machine to do

this or SAS Integration Technologies if SAS lives on a server. SAS Metadata Connections can also be set up via this dialog and data can be visually inspected prior to downloading.

Through this connection, SAS Code may be submitted. Additionally, through this connection, there are also a number of options for graphics that can easily be implemented, such as the familiar ODS graphics options. However, for the purpose of this paper, the connection will be used to create datasets and open them in JMP.

These two methods are particularly useful, as it allows JMP to be used as an analytical hub to connect to SAS. On the topic of JMP as an analytical hub, it would be remiss to not mention that similarly, scripts can be submitted to R through JMP in a similar manner. It is this author's opinion that in many instances, using JMP as an analytical hub would be ideal. JSL is object oriented and the majority of it is automatically generated from using JMP as is.

The following script uses the `sassubmit` and `sas import data` functions in JSL (JMP Scripting Language) with a SAS connection to produce a dataset, `sgf2011b` and open it within JMP.

```
sassubmit(
  "data sgf2011b (drop=i);
  do i=1 to 10000;
    v=ranuni(1900);
    w=rannor(1950);
    x=ranuni(1978);
    y=rannor(1990);
    z=rannor(1971);
  if i <5000 then c='G1';
    else c='G2';
    output;
  end;
  run;"
)
;
sas import data("work", "sgf2011b")
;
```

Finally, one can open a separate SAS Session and create the dataset of interest and write it out using a dataset library. Once this is done, this file can be opened in JMP.

Alternatively, the SAS Code below will generate a SAS dataset called `sgf2011.sas7bdat` in a SAS Session. This file will be identical to the file produced above.

```
libname ddl "C:\GJH";

data ddl.sgf2011 (drop=i);
do i=1 to 10000;
  v=ranuni(1900);
  w=rannor(1950);
  x=ranuni(1978);
  y=rannor(1990);
  z=rannor(1971);
if i <5000 then c="G1";
  else c="G2";
  output;
end;
run;
```

Of course, one must set the `libname` statement to a directory that is on your machine.

Once this file is complete, it is easy to read into JMP or JMP Pro. It must be noted that for all examples in this paper, the author will be using JMP 9 Pro. These examples provided here are generally simple in nature and use methods available in JMP and JMP Pro both, although the appearance of the interface might be slightly different across the two products.

Once you have opened JMP or JMP Pro, to import the SAS dataset created above, dd1.sgf2011.sas7bdat, go to the File Menu and then click Open to open the Open Data File Dialog. Once there, navigate to the sgf2011 file, click it, and then click open. This will open the dataset in a spreadsheet-like window.

JMP GRAPH BUILDER

As with any dataset, it is ideal to visualize the dataset first. The JMP Graph Builder is an excellent tool that offers the flexibility to visualize data in a multitude of ways by simply dragging and dropping variables. To open the JMP Graph Builder, go to the Graph menu and select "Graph Builder".

Once selected, the Graph Builder window will open. The window consists of a region where variables can be selected, a symbol representing their variable type, and an area that will contain graphics that the variables will be dragged to (See Figure 1).

The graphics region consists of eleven zones. These zones are called "X", "Y", "Group X", "Group Y", "Wrap", "Overlay", "Color", "Legend", "Shape", "Freq", and then the central data zone, with the caption "Drag and drop variables into graph zones". These regions interact with one another, depending on how variables are dragged into the graph builder.

Specifically, the graph builder is a "smart" interface in that it will attempt to build the optimal graph based on the variables chosen, their variable types, and which zone they are placed in the graphics region. For those knowledgeable in using Microsoft Excel Pivot Tables, the functionality of the Graph Builder can be considered similar, but instead of producing tables (or a somewhat dry array of Pivot Charts), the Graph Builder can produce a much richer array of different visualizations.

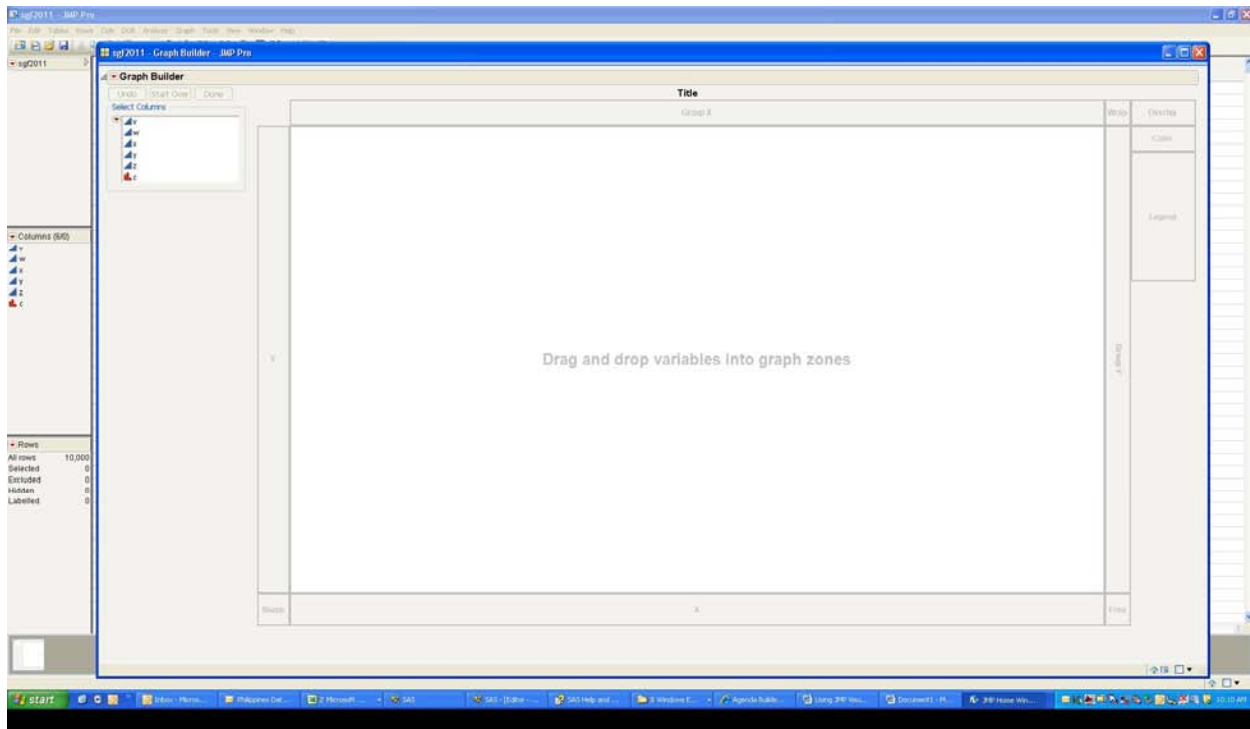


Figure1. JMP Graph Builder Main Screen

To better understand the meanings of these zones and the types of insights that can be gathered from the Graph Builder, a number of examples are provided below.

Consider dragging the variable z, or any other numeric variable onto the central zone, the x, or the y areas. This one variable plot will result in a box and whisker plot (oriented dependent on where the variable was dropped). Variable z dropped into the y area is shown in Figure 2.

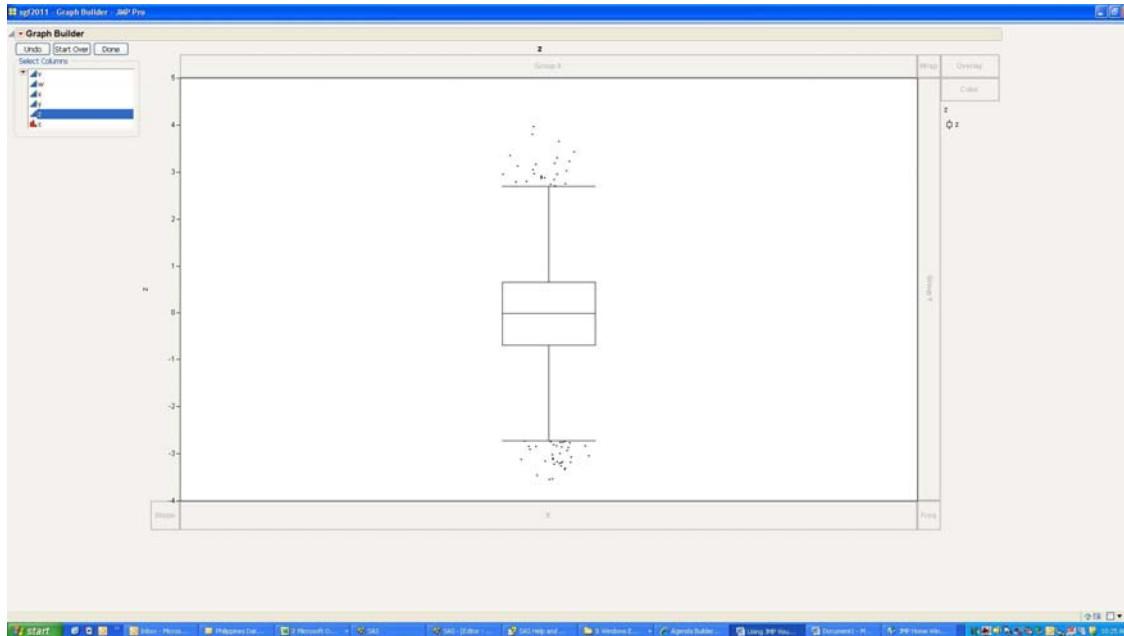


Figure 2. Box and Whisker Plot

Now, with this graph still up, variable v, another numeric variable is dropped onto the x area. A smoother graph of y versus x is produced and seen in Figure 3. To create a line graph, scatter plot, or contour plot one can right click the image and choose "Smoother" and then "Change to", and then whichever graph is desired.

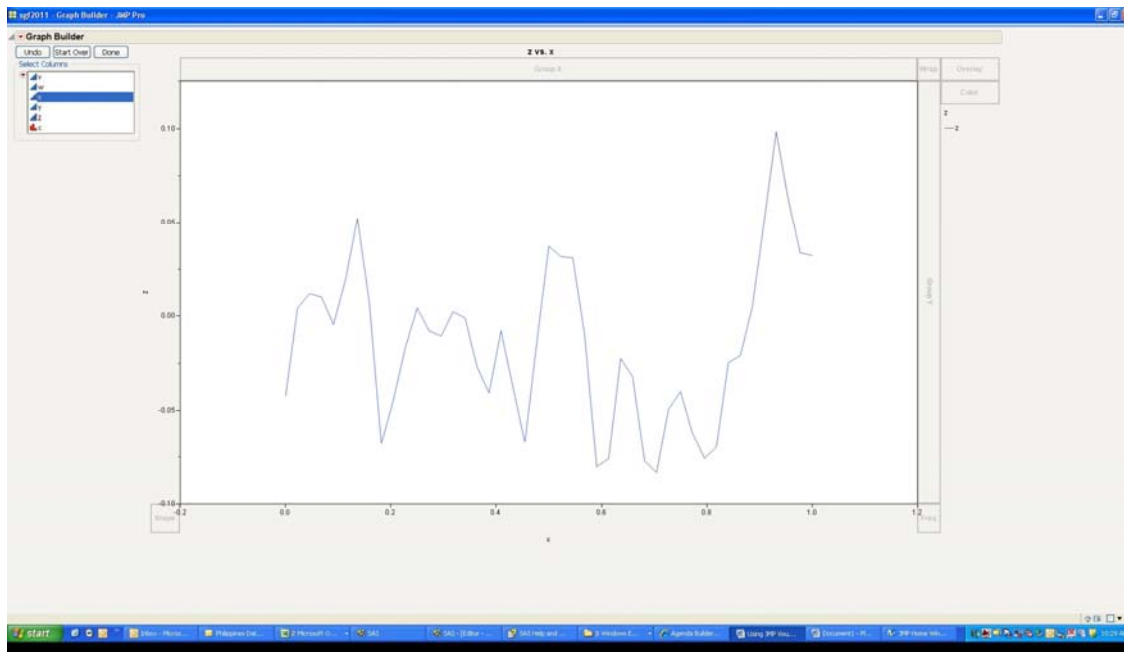


Figure 3. Smoother Graph

If variable c, the categorical variable is now added to the Group Y area, the trellis graph shown in Figure 4 is produced.

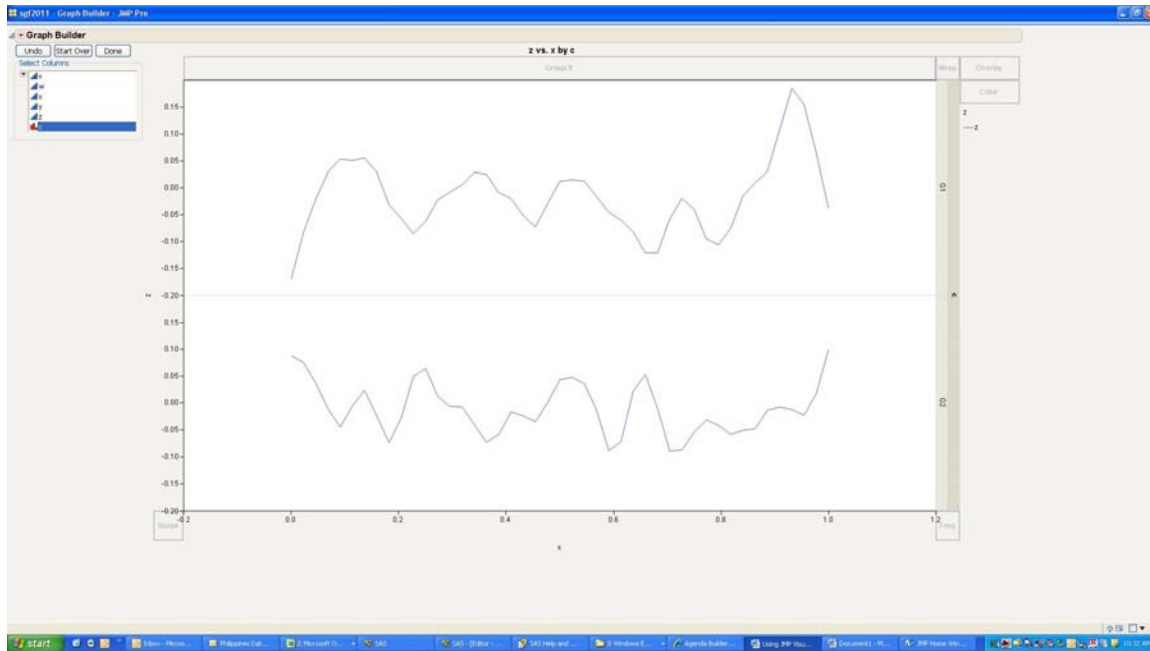


Figure 4. Trellis Graph

Now if variable y is dropped into the overlay box, the smooth lines are split into multiple lines in each rectangle of the trellis. These lines represent levels of y . In this case, since y is a continuous variable, JMP automatically creates the five y -levels seen in the graph. Typically, one might use a categorical variable in overlay, but the graph builder, being a smart tool, will attempt to produce results.

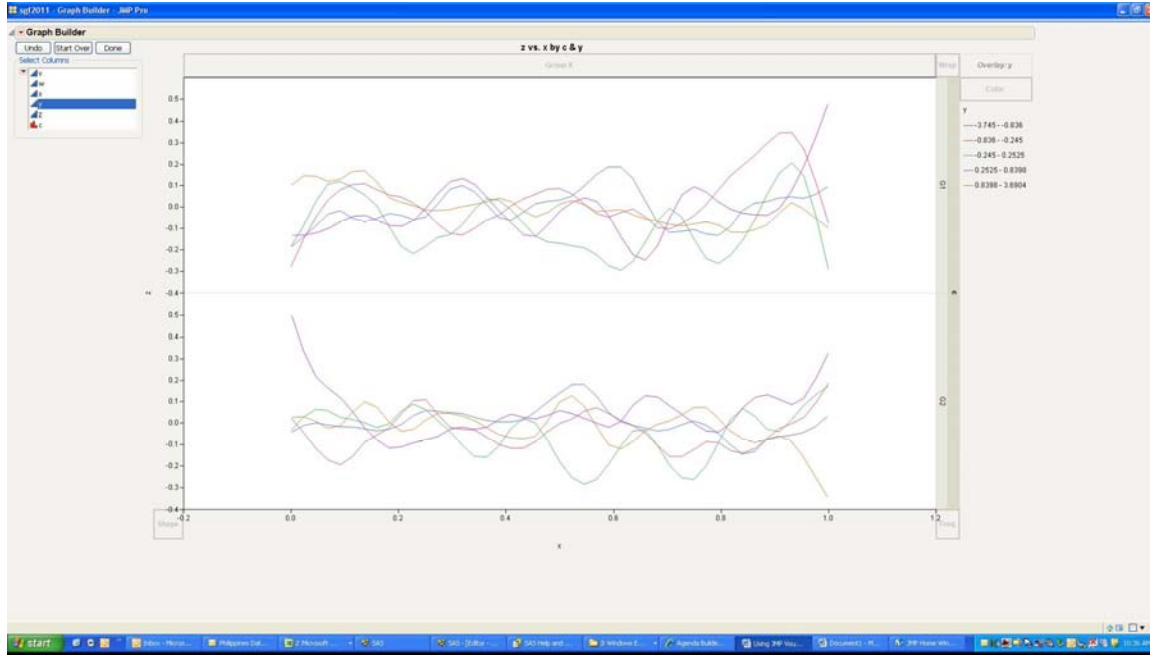


Figure 5. Trellis Graph Overlay by Levels of Y

It can be seen that with limited effort, a large number of graphs representing various visualizations of the data can be created, generating useful insights. There are multiple examples available, but the easiest way to master this interface is to open up a dataset and drag and drop variables and see the results.

For those who have coded in SAS/GRAPH®, it should be quite obvious that creation of this array of plots would take a considerable exercise in coding.

JMP DISTRIBUTION ANALYSIS

JMP offers an excellent tool visualize the distribution of your data. This can be found under the analyze menu and is quite intuitively titled “Distribution”.

Once distribution is selected, the distribution dialog box appears. This dialog allows you to select columns of interest and how they are to be used. Using the sgf2011 SAS dataset, select x and z as the Y Columns, c as the “by” column and click ok.

A window, illustrated in Figure 6, opens, showing the distribution of the data in variable x and variable x, side by side. Further, this window is separated for each level of c, the “by” column. Here, we quickly see that z is roughly normally distributed and that x is roughly uniform (as expected).

In addition, this is an interactive window. If you click on a bar of one graph, the data points that correspond to it highlight on the other graph.

While this analysis could be replicated with proc univariate in SAS, the interactivity and learnings that are available from the interactivity would not be there.

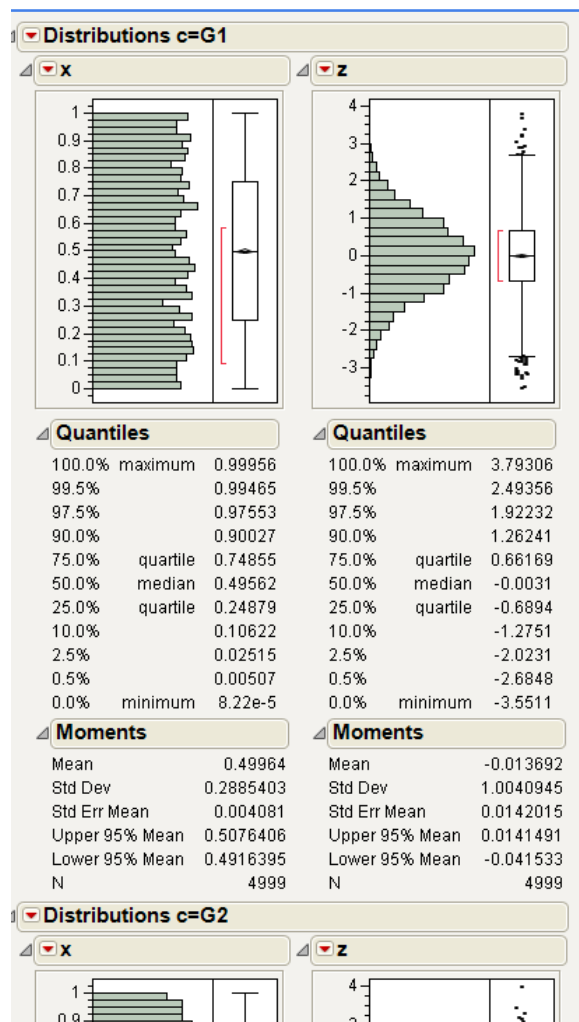


Figure 6. JMP Distribution Example

FITTING A REGRESSION MODEL

The model fitting platform in JMP provides an intuitive platform in which statistical models can be developed. JMP provides many sample datasets. To access these datasets, go to the help menu and select sample data.

Another dataset is generated for the purposes of this example. The JSL code is below.

```
sassubmit(
  "data sgf2011c (drop=i);
  do i=1 to 10000;
    x1=rannor(1978)+2;
    x2=rannor(1999)-1;
    x3=rannor(1971)+1;
    x4=rannor(2006);
    x5=rannor(1888);
    x6=rannor(1234);
    y=x1*2+x2*3+x3+x4*6+rannor(2009);
  output;
  end;
  run;"
)
;
sas import data("work", "sgf2011c")
;
```

As can be seen, this dataset is designed to setup a multiple linear regression model. This model was chosen due to its applicability to a broad audience. However, JMP handles a wide arrange of models and fitting.

Since data was created for this example, plot examples as shown above would be redundant. However, for a real dataset, one would want to use the graph builder and distribution plots to get a feel for the data and verify that it is reasonably distributed; there are no outlier issues to consider, etc. In this instance, we will proceed straight to modeling and will consider linear models. To do this, the Analyze Menu is chosen. Under Analyze, Fit Model is chosen. Here, Y is chosen as the Y variable and each of the x variables is chosen as the x variables.

The stepwise platform will be taken advantage of to five stopping rules available for model selection, BIC (Schwarz's Bayesian Information Criterion), AICc (Corrected Akaike Criterion), p-Value threshold, Max Validation RSquare, and Max K-fold RSquare. In this case, since a validation sample was not specified, the two RSquare methodologies are grayed out. Further, other options that one typically associates with stepwise regression are available, such as direction, entering all variables to start, etc. Here, AICc will be chosen with forward selection, and no variables entered. Each step in the model selection can be taken using the "Step" button, or alternatively, the "Go" button allows JMP to quickly determine the model based on the specified criterion. The model chosen by JMP is as expected, with x1, x2, x3, x4 all selected having coefficients of 1.99, 3.01, 1.00, and 5.99, respectively. Further, a very high r-square is seen. However, often, using real data, one may want to look at the effects of non-selected variables in the model. This can be accomplished by manually checking the boxes in the "Current Estimates" section of the window. Further, if there is a variable of interest that is being assessed (as in the case of a hypothesis test), it can be "locked" in, while the model is built around it. Here, we accept the model selecting x1, x2, x3, and x4, by choosing "Run Model" with those four selected. This brings up the "Fit Model" Window.

For comparison, this analysis would be done using proc reg in SAS. For more complicated datasets with a lot of exploratory analysis necessary, it would take multiple iterations of proc reg.

JMP STATISTICAL GRAPHICS AND JMP PREDICTION PROFILER

The fit model window provides very similar results by default as proc reg would. However, by right clicking each section, available results for each section can be seen. For example, by right clicking on the "Parameter Estimates" section, VIF can be displayed. Here, it is seen VIFs are nearly 1, and as expected, there is little evidence of multicollinearity.

By right clicking the red triangle near fit group, various statistical graphics can be seen, dependent on model type. In this case, surface profiler, contour profiler, and profiler are available. Profiler is the prediction profiler and will be

illustrated here. By selecting the prediction profiler, a box with a line for each variable is seen. Interactively, levels of each variable can be set, and the response changes. This mechanism is useful to understand visually how each variable affects the response. Figure 7 illustrates this.

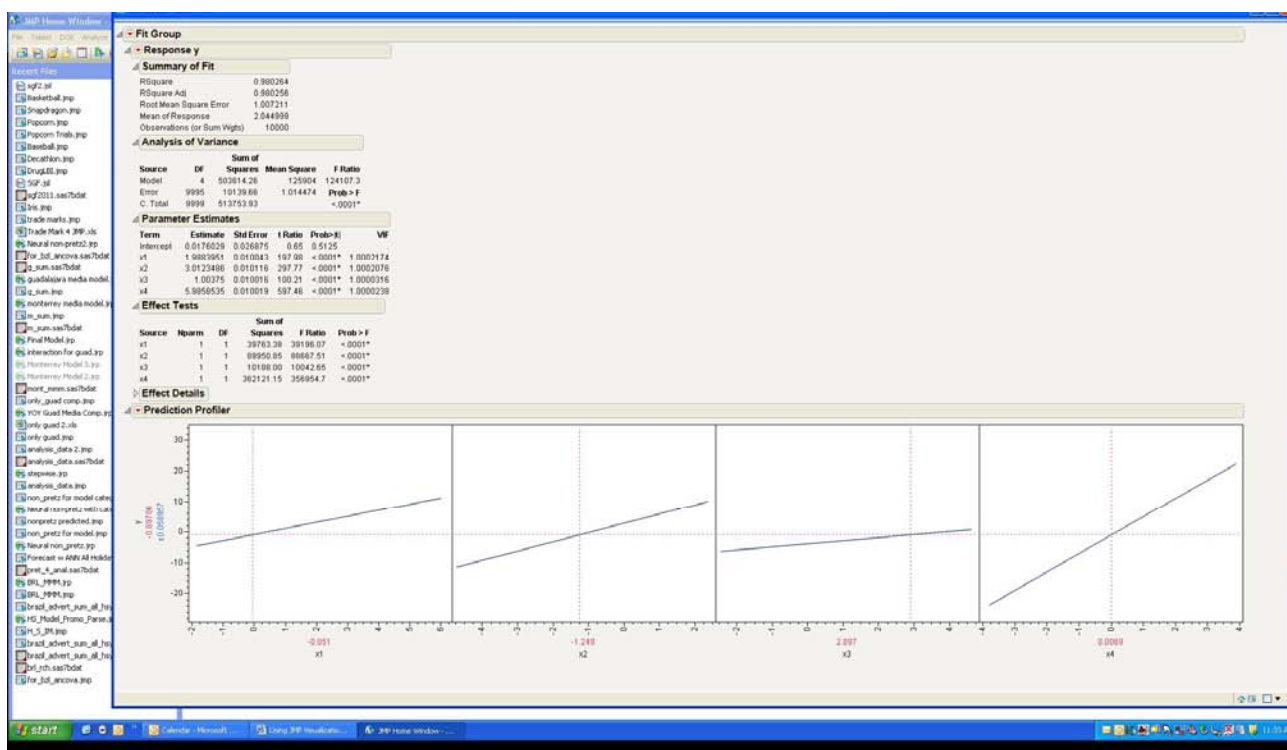


Figure 7. JMP Prediction Profiler

Further graphics options can be seen by right clicking the red triangle next to “Response y” and browsing through the options, depending on the graph desired. By selecting “Factor Profiling” for instance, “Cube Plots” can be generated. However, it is seen that “Interaction Plots” is a grayed out (under “Factor Profiling”). This is because there is no interaction term in the model.

JMP INTERACTION PROFILER

Now consider an ANCOVA model with an interaction term. The following code creates a dataset that would be reasonable to model in this manner.

```
sassubmit(
  "data sgf2011d (drop=i);
  do i=1 to 10000;
    x1=rannor(1978)+2;
    x2=rannor(1999)-1;
    x3=rannor(1971)+1;
    x4=rannor(2006);
    x5=rannor(1888);
    x6=rannor(1234);
    if i > 5000 then x7='a';
    else x7='b';
    if i > 5000 then y=x1*2+x2*3+x3+x4*6+x1+rannor(2009)+5*rannor(2100);
    else y=x1*2+x2*3+x3+x4*6-x1+rannor(2009);
  end;
  output;
  run;"
)
```



```

;
sas import data("work", "sgf2011d")
;

```

Since the functional form of this model is known (and the stepwise functionality of JMP has been illustrated), the model $Y=X1+X2+X3+X4+X7+X1*X7$ is selected within "Fit Model", but with personality "Standard Least Squares".

Graphically, for this model, the prediction profiler and interaction plots are selected (See Figure 8). The interaction plots are useful, as it becomes clear how the level of variable x7 affects the slope of variable x1. This type of relationship is very important in a variety of fields, ranging from marketing to pharmaceuticals.

Furthermore, the prediction profiler is still available, but of interest is that as the levels of x1 or x7 are toggled, as expected, the slope of the other line changes, which is by definition, the meaning of an interaction term.

Using the interaction plots and prediction profiler is particularly useful in these instances when the models are more complicated. The models can be easily saved to the table and automatically updated for new data coming in.

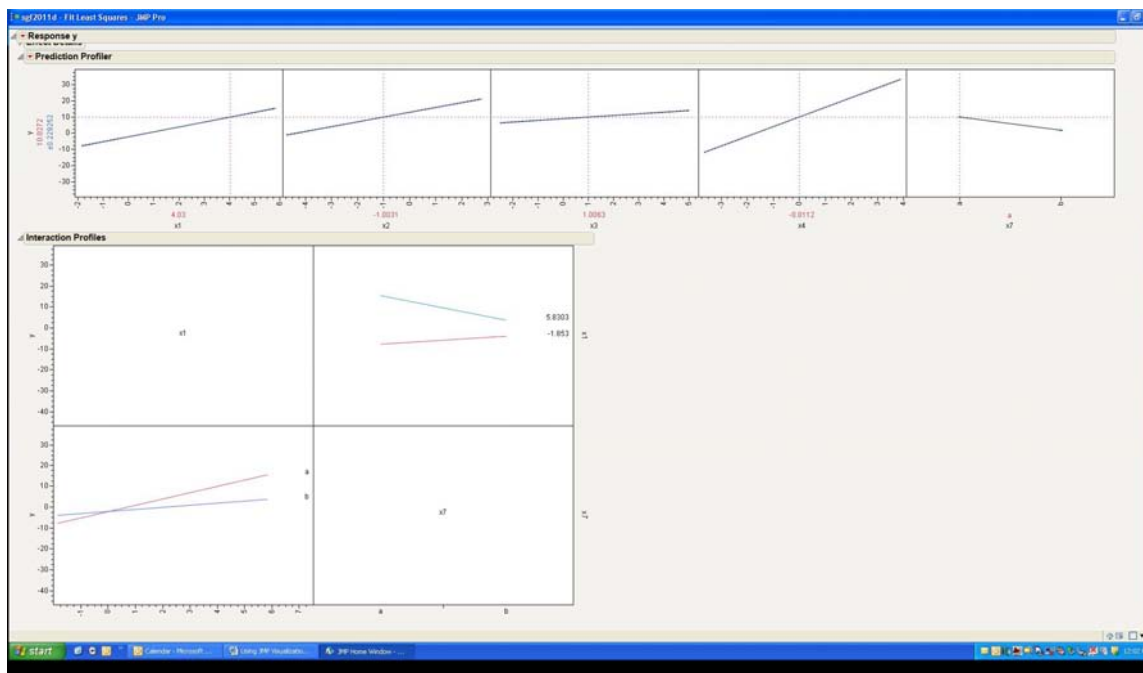


Figure 8. Interaction Example

JMP NEURAL NETWORK INTERFACE

Finally, a quick illustration of the Neural Net profiler will be shown, using this same simple dataset. Some of this functionality may only be available in JMP Pro.

To access the Neural Network Interface, from menu, select "Analyze" and then select "Modeling" and then "Neural". A so-called "Kitchen Sink" model is selected here for the purposes of illustration (with x1-x7 the x variables and y as the y-variable). Once done, select ok (note since this is for the purpose of illustrating a particular graphic, validation data sets, and other complex methodologies will not be discussed.). On the next menu (for the same reasons) all defaults will be selected. By right clicking the triangle next to "Model NTanH(3)", the profiler can be opened, giving the same type of graph as seen in the linear model above. Like above, the relationship between the variables the ANN (Artificial Neural Network) created can easily be visualized. Further if "Diagram" is also selected a diagram, illustrating the relationship between the variables and the nodes, is produced. Figure 9 shows these visualizations.

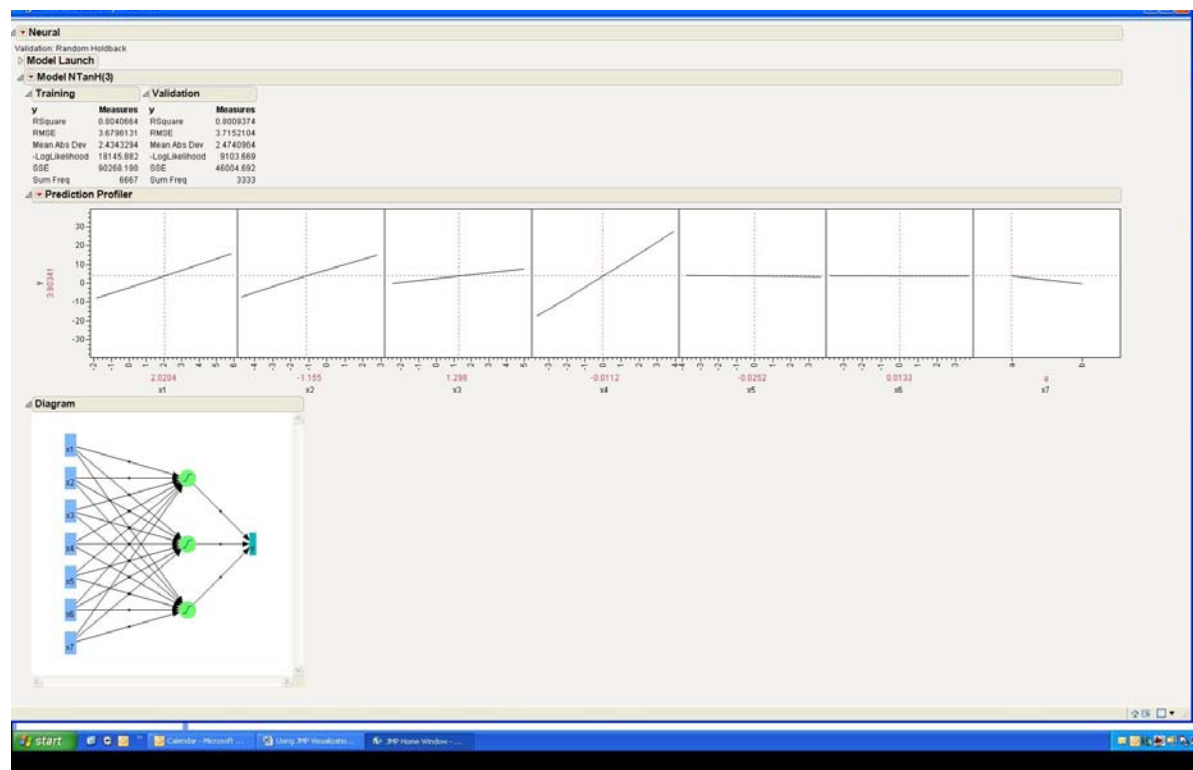


Figure 9. ANN Graphs

In the case of ANNs, this profiler is particularly useful, as they are often “black boxes” with respect to predictions. That is, while their predictive ability is often quite good, the interpretation of the model is often not well understood. While this example is quite simple, these graphs are available for any ANN that can be created within JMP.

JMP can also generate a SAS code at this point by choosing the red hot spot and selecting Make SAS Data Step. This allows experienced SAS users to leverage JMP as a first round modeling tool and further augment/tweak their models in SAS.

CONCLUSION

JMP is a powerful tool that provides a wide array of statistical visualizations. This paper demonstrated some of the more useful visualizations, but only scratched the surface as to what visually JMP can produce. It is believed that these visualizations can help most statisticians and analysts better understand their data and produce new and useful insights.

ACKNOWLEDGMENTS

I would like to acknowledge the assistance of Aashish Majethia of JMP for proofreading this paper, making some great suggestions and for showing me how to submit SAS Code using JSL. Further, I would like to thank my manager, Rafael Alcaraz, Ph.D., and The Hershey Company for their support. Finally, I'd like to thank my wife, Debbie, who is putting up with me finishing this paper off on a day I'm supposed to be taking vacation.

RECOMMENDED READING

Read through the JMP help files. They are incredibly readable and often have useful examples.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

George J. Hurley
The Hershey Company
19 East Chocolate Ave.
Hershey, PA 17033
717-534-5337
ghurley@hersheys.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are trademarks of their respective companies.