



## Inside This Newsletter

Opportunity Space at  
Georgia Tech Aerospace  
System Design Labs 1

Modeling Time associated  
with Business Processes 4

A Script Quickie for List  
Users 9

Copy and Paste to Assign  
Value Labels 9

Obtain and Interpret  
Odds Ratios for  
Interaction Terms 10

Upcoming Conferences 13

Do You Know About  
Trainer's Kits? 12

Spline Models 14



*JMP desktop statistical discovery software from SAS uses a structured, problem-centered approach for exploring and analyzing data on Windows, Macintosh, and Linux. The intelligent interface guides users to the right analyses. JMP automatically displays graphs with statistics, enabling users to visualize and uncover data patterns.*

## A Flight Through Opportunity Space at Georgia Tech Aerospace System Design Labs

*By Diana Levey, SAS Institute*

How can Congress plan a budget capable of taking us back to the moon, Mars, and beyond? How can aerospace companies find their way through the technical frontier to design a supersonic business jet? How does the U.S. Air Force get the least drag, the most speed, and the optimal range in the next generation of fighter jets?

Georgia Institute of Technology's Aerospace Systems Design Laboratory (ASDL) enables these dreams to come true. With end goals of improved safety, reduced environmental impact, and lower acquisition and operating costs, Director Dimitri Mavris, Michelle Kirby, and the other research engineers use state-of-the-art JMP software in a one-of-a-kind laboratory setting. There they help organizations from industry, the military, and government develop the world's most sophisticated engineering designs.

Most of the projects are related to aerospace, and all are complex. Kirby says, "The systems designs are so inter-dependent that collaborative planning becomes more valuable than in other engineering disciplines." Mavris says, "This is the lab to do that work. ASDL is the only group that is this advanced in systems-of-systems thinking. We are at the cutting edge right now."

Mavris explains that before ASDL, universities were producing graduates who lacked some of the qualifications industry leaders sought when hiring employees. Mavris set out to close that gap, launching ASDL in 1992 to link education and industry, detailed design and manufacturing, and science and engineering. "Students come in as propulsion or fluid specialists, and leave as more valuable systems integration specialists," says Mavris. Despite the fact that government and industry spend some \$8 million at the lab in a given year, Mavris says: "We're not here to make money; we're here to create the next generation of engineers."

While Mavris and his 25 staff members groom undergraduate and graduate students for real-world success, the 200 students themselves help create the next generation of affordable and high-quality complex systems for industry and government.

### Adding the Genius of JMP

In 1994, Mavris discovered JMP statistical discovery software from SAS and decided to add it to his laboratory's tools. When he saw JMP demonstrated in a life sciences environment, Mavris says, he knew the software would enable his team to conduct true experiments. He recalls, "I watched a demo with mice and I saw rockets." Right away, he knew JMP would allow the interactive analysis and data visualization needed to figure out the best scenarios for such complex systems. "The visualization ... whoever discovered that visualization is a genius," Mavris says.

"We don't want homegrown tools anymore," he says. "We want a complete package, and that's what JMP is. I consider it an empowering tool, a tool for the practitioner. It empowers you to do 100 times the work you could do." Mavris said he tried Minitab, but, "we assessed it and it wasn't standing up." Now, in addition to JMP's use in the lab, JMP is used by nearly every student that comes through the master's program.

## A Collaborative Visualization Environment

But just using JMP in the traditional fashion wasn't enough. "This interactive statistical software warranted a special environment," says Mavris. "We wanted to create an entity like no other." And that he did. The resulting lab is called the Collaborative Visualization Environment, or CoVE, and it features a 10-foot tall by 18-foot wide multimedia wall made out of 12 screens. "It creates a mosaic of information—a place to showcase all of this data."



Why so big? Because the number of variables is huge, as are the models. And the huge display enables research partners to clearly see and understand the systems models. Prior to building the CoVE, Mavris would show JMP profiler displays that would have to be scrolled, which would result in a less effective presentation. "After you scroll up and down a few times, you lose your audience," he says.

"But not in the CoVE," says Mavris. "After they see it, they get it. For every product, the breakthrough has almost always been in JMP."

## Exploring Options Interactively

Some of the physics-based computational dynamic models, like those from NASA, are expensive to run and simply cannot be run in real time for what-if analysis. Yet hands-on decision making depends upon real-time modeling. Mavris' challenge: How can his team explore options interactively, when it takes hours to run just one design past the physics models? The answer: surrogate modeling. Like the name implies,

surrogate statistical models can represent expensive physical models, proprietary mathematical models, or computer-intensive simulations.

The surrogate models actually become the communication tool for the huge, sometimes secret, and often competitive types of projects ASDL takes on. If the engineers worked with all the real parameters, then vendors or groups of research partners would have to share proprietary information. Surrogate models are a safe way to share information without having to worry about competitive concerns. These models cannot be reverse-engineered, so it's how political and organizational barriers are handled.

"Imagine working on something that won't go into production for a number of years. Engineers shouldn't be constrained by current parameters, which will be outdated by the beginning of the production process. Instead, the partners can use variables rather than constants. All the plans can be carried out hypothetically and conditionally," Mavris says.

"And with such huge amounts of data and so many variables, surrogates allow us to look at information in bulk instead of focusing on details," Mavris says. "They actually speed up the process." Mathematical models of this scale are too big to easily and quickly conduct what-if scenarios. So if surrogate ones are done first, then the expensive, mathematical one that follows should be accurate. The latter is run on a Dell supercomputer system that, when combined with a similar system for the Georgia Tech physics department, provides 1,024 processors running in parallel.

## Filtered Monte Carlo in Action

Another important way of modeling at ASDL: Filtered Monte Carlo. Graduate researcher Pat Biltgen demonstrated how JMP is used in a military target simulation involving Air Force strike aircraft and anti-aircraft defenses. The goal is to see what weapons system will penetrate to a target, and what the characteristics of successful systems are. The major technical challenge in this research is the need to quickly execute probabilistic analyses and visualize complex, multidimensional results.

Using surrogate models, point designs can be quickly generated using accurate approximates of physics-based design tools. Surrogate models also enable rapid Monte Carlo simulations to be run nearly instantaneously. These two techniques are combined to enable capability-based design and technology exploration. Using uniform distributions on the subsystem-level input parameters, the effectiveness of a proposed

solution at the 'system-of-systems' level can be evaluated. The Filtered Monte Carlo method is then used to 'filter' or reduce the number of solutions from hundreds or thousands to a handful of points by applying constraints at the top level and identifying solutions left at the system and subsystem level.

Now the job has changed from exploring a complex, multidimensional mathematical space into querying a database of simulations very rapidly. Even with hundreds of thousands of points, engineers can show all the possibilities graphically in many directions with scatterplot matrices. Then they start brushing points, and dragging over areas that they want to select or infeasible points that they want to exclude until they see points defining the opportunity space that satisfies what they are looking for.

Whereas profiling only shows up to three dimensions at a time, Filtered Monte Carlo shows opportunities in many more dimensions. Each point is a probe, and each surviving point that is not excluded by the conditioning remains to define further options. And that's what the engineers are after: options. A mathematical optimization will find just one solution, but a Filtered Monte Carlo will find all the combinations that work. It will present choices. After all, there are many ways to hit the target.

Another technique used in this exercise is the surface profiling feature, which generates three-dimensional 'cubes' that exercise the parametric surfaces captured by the surrogate models. Biltgen, describing the 64 discrete behaviors the team can show in the Air Force simulation, said, "We could have eight cubes and all of those cubes start break dancing as the parametric slide bars are moved." Biltgen's description: "This is an awesome capability!" According to Biltgen, the ability to simultaneously query complex multi-dimensional spaces in a graphical manner helps enable capability-based design. "The primary problem that we have is that we can generate hundreds of thousands of designs, but no one is ready to see hundreds of thousands of designs," says Biltgen. "The visualization capabilities of JMP bring the decision maker into the process and allow him or her to see large amounts of information that cannot be explained in any other way."

## Next Generation of Business Jets

"Awesome" might also describe the next generation of business jets—another ASDL project. Here, engineers analyze performance tradeoffs in the early phases of design, allowing consideration of variables involving noise-level requirements, environmental emission standards and fuel consumption goals. "The methodology maps propulsion characteristics to overall

system metrics such that the entire design space can efficiently be examined," writes Simon Briceno and Mavris in *Quiet Supersonic Jet Engine Performance Tradeoff Analysis Using a Response Surface Methodology Approach*. "The design essentially has an analytical means to examine every conceivable alternative within the design space."

The result: a quiet supersonic jet engine that meets all regulations and costs less to operate. "They allow the decision maker to play what-if games and make tradeoffs in design early, knowing the system-level consequences of those tradeoffs. The designer is given an understanding of the magnitude of impacts that different design parameters can have on responses."

In a paper written for a World Aviation Conference, ASDL's Peter Hollingsworth and Mavris look at a practical what-if scenario for the Hypersonic Strike Fighter, a fighter jet that can exceed five times the speed of sound. And, they conducted the concept exploration in the presence of open and evolving requirements. One what-if question: What if they wanted to change a hypersonic vehicle from a land-based aircraft to a carrier-compatible system?

"Most likely," states the paper, "the designers would never have considered implementing the modularity or technologies necessary to achieve carrier compatibility in the initial vehicle design. Because of this, the addition of the carrier compatibility requirement (restraint) may render the system infeasible."

Rendering hypersonic fighters, supersonic jets or any other complex system infeasible is not an option anybody wants to consider. Kirby says that it's better to weigh all of the options in the early stages of design, when what-if scenarios can be played out and the many tradeoffs considered. "JMP is at the core of this; it's been key," says Kirby. "After all, we are JMP power users."

Learn more about this premier center for the development and application of advanced design methods for complex systems at

[www.asdl.gatech.edu](http://www.asdl.gatech.edu)

*The Aerospace Systems Design Laboratory was founded in 1992 and has grown to be one of the nation's premier centers for the development and application of advanced design methods for complex systems, and the training of the next generation of engineers and scientists.*



# Modeling Time Associated with Business Processes

By Mark Bailey, SAS Statistical Training and Technical Services

A common business problem is reducing the time that it takes to finish activities such as handling customer calls, processing orders or payments, bank transactions, or even waiting on patrons at a restaurant. A significant reduction in time translates into lower costs, more available resources, and higher customer satisfaction. These kinds of problems are often the focus of Six Sigma projects.

A Six Sigma team first collects data to understand the nature of the problem and again after a proposed solution is implemented to see if a time reduction actually occurred. Many situations only consider using a normal distribution model or an individual measurement control chart. These basic tools can reveal much about the activity. Other analytical tools are available that give more informative about the nature of the problem and the result of change in procedures.

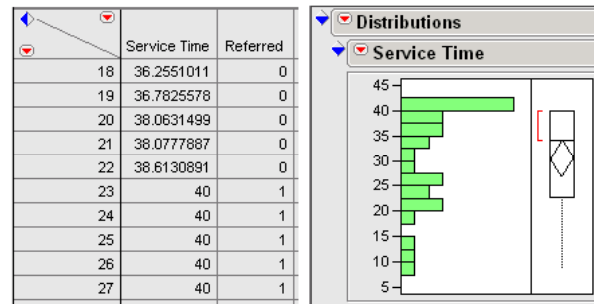
This kind of data is referred to as *time to event* or *event time*, where event means closure such as finished a service call, processed an order, payment, or transaction, or served a meal. The analytical methods for this kind of analysis were originally developed for the events involving mortality or machine failure. These methods are referred to as *survival* or *reliability* analysis. (Allison 1995) This analysis has its own established conventions and terms, which will be introduced here. More advanced theory and applications are available from other sources. (Nelson 1982, Meeker and Escobar 1998).

Event time data has several notable characteristics. It is usually not symmetrically distributed so confidence intervals are not symmetric and the mean and median are not usually equal. The data are not modeled well by the normal distribution. Instead, three other distributions emerge as the most useful: Weibull, lognormal, and exponential.

## An Example

The data table called **Sample1.jmp** lists the time, in hours, to complete a customer request by the technical service department of a local cable television company. The data tracks new requests from the start of the job either to completion or referral to a specialist after the allotted 40 hours. *Figure 1* shows an example of the data and the result of Distribution platform. The distribution shows clearly that the data is not normally distributed.

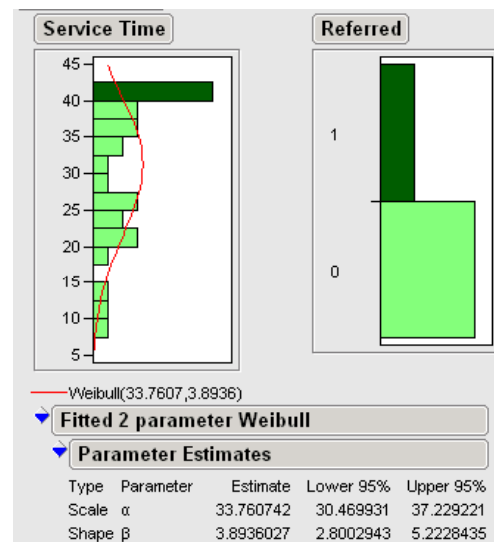
Figure 1 Partial Listing of Sample1 Data and Distribution of Service Time



The results of the Distribution platform in *Figure 2* show a Weibull distribution fit, and also reveal a new problem. Some of the requests were not completed in the allotted time. The time recorded in these cases was the last time known (40 hours). The actual completion time is greater than 40 hours and isn't known. This phenomenon is referred to as *censoring*. In some cases a significant portion of the data are censored values. JMP offers analytical methods specifically designed to handle censored data.

You could delete the censored data before running this analysis but with a loss of information that could bias the analysis. In some cases a significant portion of the data are censored values. A better way is to use methods that handle censored data.

Figure 2 Histograms Show Nonnormality and Censored Data





## Survival/Reliability Analysis

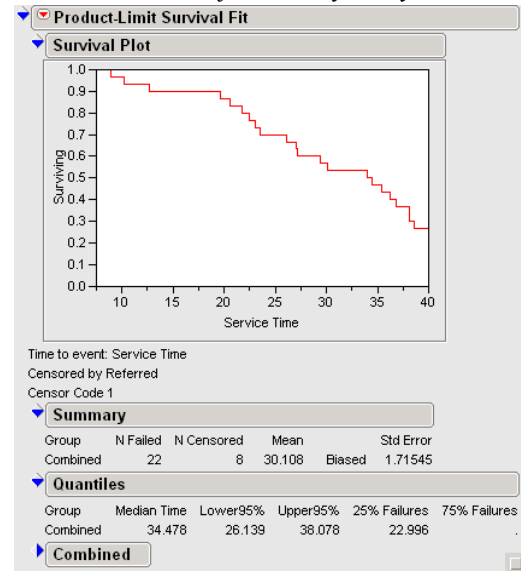
The column called **Referred** in the **Sample1** data table shows 0 if the work was completed and 1 if it was incomplete and referred to a specialist for further action (censored). You can analyze the data with the Survival/Reliability platform in the Analyze menu. This platform recognizes censoring and provides an analysis role in the launch dialog for this purpose. The new column, **Referred**, is cast in this role as shown in *Figure3* (below).

The initial report, shown in *Figure 4*, includes a survival plot and general descriptive statistics based on the Kaplan-Meier product-limit model. The plot shows that all requests are initially open (Surviving). The proportion of requests surviving (still open) declines as time increases. Note in an actual reliability or survival study, the hope is to extend the time. You are using the same analysis methods here but your goal is to reduce the time instead. For example, at 25 hours, 0.7 of the requests are still open (30% of the requests are complete). At the end of the allotted time (40 hours), there are still about 27% of the requests unfulfilled.

The statistics from this sample include the number closed (22 **N Failed**), the number remaining (8 **N Censored**), the mean time to closure (30.108 hours), and the median (34.478 hours). The number of closed requests is called **N Failed** in the survival report because this method is often used to look at failure rates.

The Survival/Reliability platform can perform a Weibull analysis that accounts for censoring. Select Weibull Plot and Weibull Fit from the platform menu to see the plot in *Figure 5* with the axes transformed so that the data points form a line if the Weibull distribution is a good model. Now the x-axis becomes  $\log(\text{time})$  and the y-axis becomes  $\log(-\log(\text{survival}))$ . The right side of the plot shows a probability scale. As time increases (left to right), the probability of survival (service request remains open) decreases. Note that the

Figure 4 Initial Results of Reliability Analysis



probability scale is inverted. As the  $y$  values increase, the probability goes down. The fit agrees well with the data. Only a couple of points in the lower left corner are not close to the fitted line. The crosshairs tool is useful for reading coordinates along the fitted line. Try using them to find the median service time—the time when 0.5 or 50% of the calls are resolved.

Figure 5 Weibull Fit and Weibull Plot

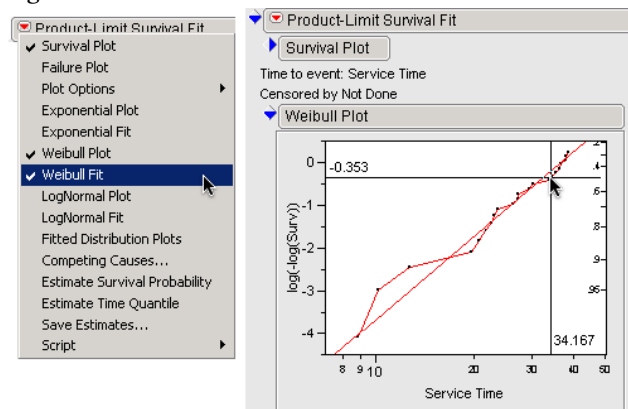
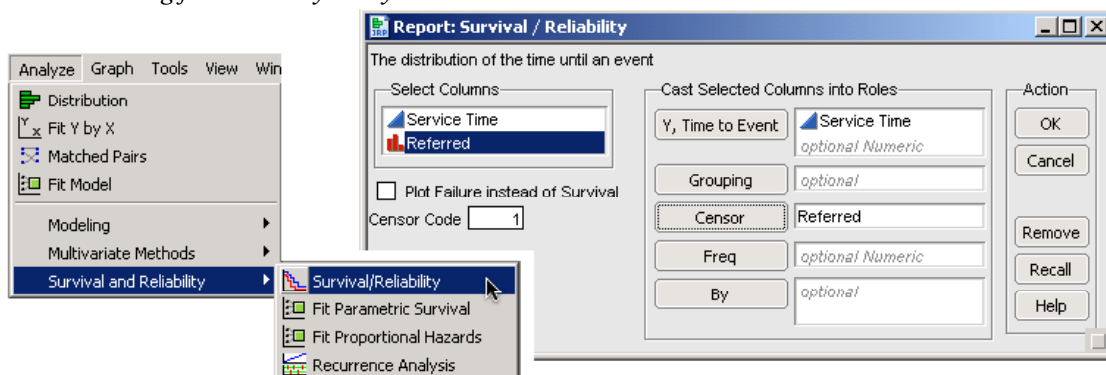


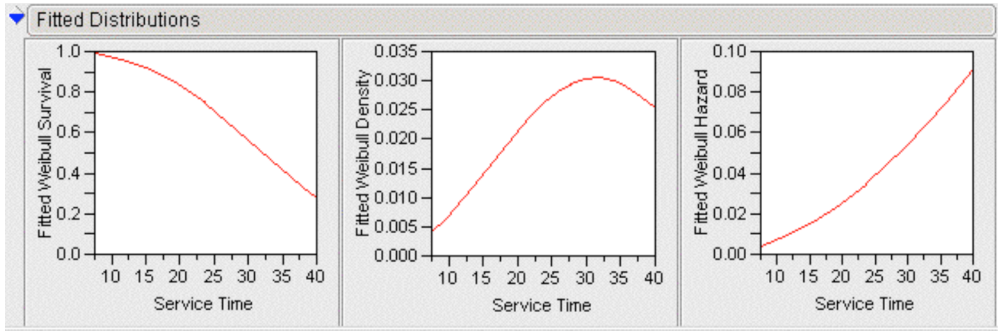
Figure 3 Launch Dialog for Reliability Analysis



The fit is reasonable. What can we learn from it? Select the **Fitted Distribution Plots** command from the platform menu to see the plots in *Figure 6*. The plot on the left shows the survival (open requests) versus time. The plot in the middle is the skewed probability

density function you expect, not a bell-shaped normal distribution. The plot on the right shows the hazard function, or the rate of failure versus time. In the context of service requests, it shows that the rate of completing requests increases over time.

Figure 6 Fitted Distribution Plots



Significant Factors that Cause Delays

Your team now considers reasons that prevent handling a service request in less time. A list of four causes is adopted for the study.

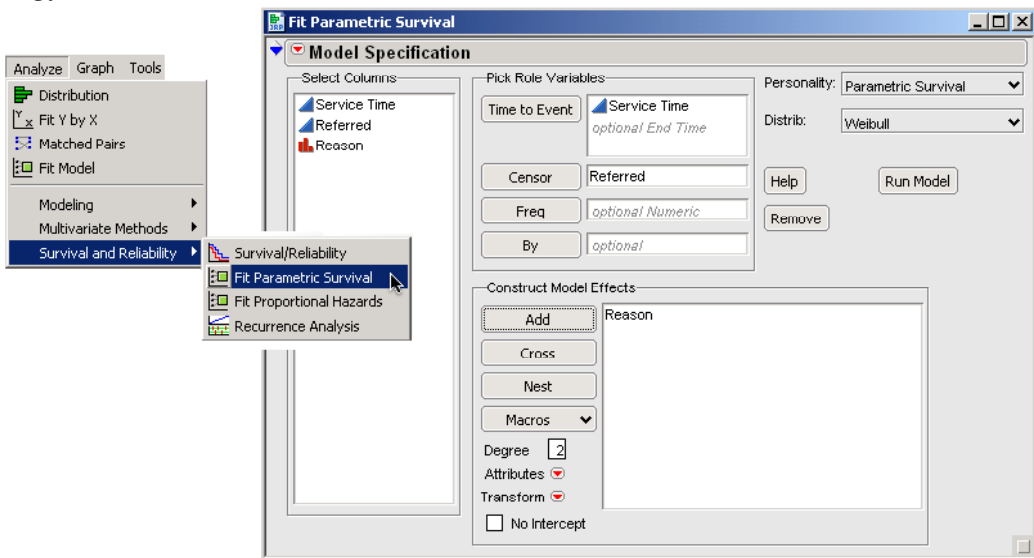
- incorrect information (“Wrong Info”)
- traffic or weather conditions (“Traffic, Weather”)
- customer not home (“Not Home”)
- tools or parts unavailable (No Parts)

The data is in the **Sample 2** data table (*Figure 7*). The third column shows the reason for the delay in completing each request. A new analysis can show how these reasons affect the distribution of completion times. Select **Survival/Reliability > Fit Parametric Survival** in the Analyze menu (*Figure 8*).

Figure 7 Partial Listing of the Sample 2 Data Table

Sample2				
		Service Time	Referred	Reason
Survival				
Parametric Survival				
Proportional Hazard				
Columns (3/0)				
Service Time				
Referred				
Reason				
Rows				
All rows	38			
Selected	0			
Excluded	0			
Hidden	0			
Labelled	0			
		1	6.01021764	0 No Parts
		2	9.53838021	0 Traffic, Weather
		3	10.9895211	0 Traffic, Weather
		4	13.955661	0 Traffic, Weather
		5	15.2738511	0 No Parts
		6	15.7149456	0 Not Home
		7	20.2922141	0 No Parts
		8	21.2930148	0 Traffic, Weather
		9	21.6585347	0 No Parts
		10	21.8261804	0 Not Home
		11	22.5957797	0 Traffic, Weather
		12	22.8030735	0 Not Home
		13	24.7574748	0 Traffic, Weather
		14	25.2901827	0 Wrong Info

Figure 8 Dialog for Weibull Parametric Survival Fit

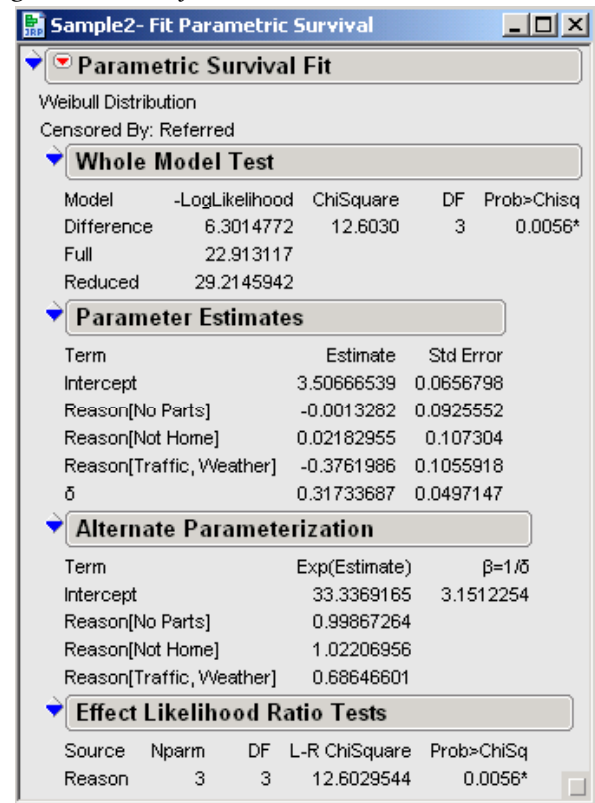


**Reason** is added as an effect in the model of Service Time. Note the Weibull distribution is still the basic model. Click **Run Model** to see the results in *Figure 9*.

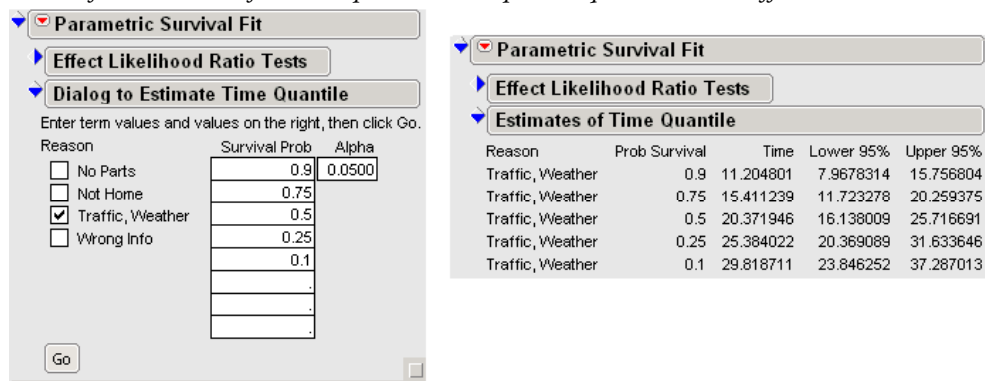
The Whole Model Test and the Effect Likelihood Ratio Tests both **Prob>Chisq** is 0.0056 because there is only one effect, **Reason**. The levels of Reason affect the Weibull distribution of the completion times. In particular, they combine with the Intercept to make the distribution scale parameter, also known as  $\alpha$ , as shown in the Alternate Parameterization report. The shape parameter, also known as  $\beta$ , is independent of the Reason levels.

The model can be used to predict the time it will take to complete a specific portion of the requests when one or more reasons are involved. For example, suppose you want to predict the time for 90%, 75%, 50%, 25% and 10% requests to remain open when Reason is "Traffic, Weather". To do this, select the **Estimate Time Quantile** command in the platform menu title bar. Complete the values as shown on the left in *Figure10*, then click **GO** to see the results on the right. The model predicts 11.2, 15.4, 20.4, 25.4, and 29.8 hours, respectively. Note confidence intervals for these time estimates are provided as an indication of the uncertainty from this sample.

*Figure 9 Results of Weibull Parametric Survival Fit*



*Figure10 Prediction of the Amount of Time Expected to Complete Request in Bad Traffic or Weather*



## Conclusion (It's Only the Beginning)

Other survival methods are available to explore this kind of data but they are beyond the scope of this introductory article. One alternative is to fit a proportional hazards or *semi-parametric* model. The effects in the model are parametric but no distribution is required, only the assumption that all cases share the same unspecified baseline hazard function. You set it up and use it similar to the way you used the parametric survival model above.

In some situations, you may want to compare groups. For example, the service requests might be handled by

several regional customer centers. Alternatively, you might track different kinds of requests for service such as new installation, equipment repair, billing inquiries, and canceling service. All that is required is a data column that records to which group an observation belongs. In the analysis, cast this column in the Grouping role. In either case, JMP computes separate and combined survival curves and tests for significant differences in survival between groups.

What if your team eliminated one or more of these reasons? Which one makes the most difference in the mean time to complete the service? How much

difference does it make? Your team can use this analysis to validate their decision on a project plan. You can run an experiment to verify the improvement. Collect another sample but now designate the status (before, after) of these cases instead of the reason for the delay. Use the same analysis as the previous example to see how the groups are different in the survival plot and if the difference is significant.

Finally, some cases close for another reason. Essentially the expected event never occurs. Perhaps a customer calls to request service but then cancels the request before it is complete because one of several things might happen: the problem goes away or the customer cancels service altogether. As you can see, these other events preclude the original event under study from happening. The set of all possible events is said to be *self-censoring*. Such a set of events can be studied by an extension to the Kaplan-Meier product-limit method

known as *competing causes*. You initiate this further analysis with a command by the same name from the Survival platform menu.

Imagine other applications of these methods that involve the time to some event: time to pay an insurance claim, time waiting to see the doctor, time to check out at the supermarket, time to commute to work or return home, time to load a truck for delivery, and so on. There are many applications for survival methods beyond their original purpose.

#### References

Allison, P. (1995), *Survival Analysis Using the SAS System: A Practical Guide*, Cary, NC: SAS Institute, Inc.

Nelson, W. (1982), *Applied Life Data Analysis*, New York: John Wiley & Sons.

Meeker, W. Q. and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: John Wiley & Sons.

### Design Institute for Six Sigma at SAS Institute

SAS offers the Design Institute for Six Sigma, with extensive Six Sigma training, superior consulting and experienced mentoring services for multiple levels of your business in both transactional and production-based organizations. Training can be conducted at SAS public training centers in the U.S. or at your business location, when it is convenient for you. For training at your location, our Master Black Belt team can develop customized Six Sigma content, including Lean concepts as well as provide mentoring and advisement services for your customized Certification development strategy.

Look to the Design Institute for Six Sigma at SAS to enable you to meet your Six Sigma goals.

#### Consulting and Mentoring Services:

Our team of Six Sigma experts can provide an end-to-end solution from the beginning of your Six Sigma project through to deployment. This includes consulting and mentoring services that can be in addition to or independent of our training services.

If after attending a course, you need additional support, our Six Sigma instructors can provide technical advice to help you apply what you learned to your unique business environment. And, for those who need hands-on assistance with Six Sigma projects, our staff is available to conduct your Six Sigma project or to lead your project team in the implementation of Six Sigma methodologies or to help you overcome obstacles.

For information about The Six Sigma course schedule, and Six Sigma consulting and mentoring services, go to:

<http://support.sas.com/training/us/css/>



# TIPS TECHNIQUES

## A Script Quickie for List Users

JSL doesn't have character arrays, and lists can be slow if you have to traverse them frequently. In particular, adding to the end of a list is slow if the list is large. But you don't have to always add to the end of a list to get it there. Instead, you can insert at the beginning, and then reverse the list.

Below is an example benchmark test to insert the numbers 1 to 50,000 into a list. The first version inserts each number at the beginning of the list and reverses the list after all numbers have been added. The second version inserts each number ( $i$ ) into the  $i^{\text{th}}$  position of the list. The first version runs far faster than the second. The difference in run time depends on the machine and operating system you use.

```
tbase = TickSeconds();
list = {};
for(i=0,i<50000,i++,InsertInto(list,i,1));
ReverseInto(list);
t1 = TickSeconds()-tbase;
tbase = TickSeconds();
list = {};
for(i=0,i<50000,i++,InsertInto(list,i));
t2 = TickSeconds()-tbase-t1;
show(t1,t2);
```

t1:0.1166666666666788  
t2:17.0333333333333

## Reminder

### JMP User Conference

The Third Annual JMP User Conference will take place June 20-21 in Cary, NC. Attend exciting and insightful sessions on topics such as Design of Experiments, Process Improvement Methodologies, Issues in the Pharmaceutical Industry, and Issues in Service and Transactional Industries. Check out the new additions to this year's conference program, which include special events such as Roundtable discussions, a Scripting Workshop, a Genomics Discovery event and exclusive new training courses.

[www.jmp.com/juc06](http://www.jmp.com/juc06)

## Copy and Paste to Assign Value Labels

Suppose that you have both a code column and the full description of the codes. You want to use the descriptions as value labels—without working too hard to manually enter the text. Here are steps to do that:

1. Use **Tables > Summary** to summarize codes and descriptions. Verify that each code has just one description.

	MHT	Habitat Type
1	1	Tropical Moist Broadleaf Forests
2	2	Tropical Dry Broadleaf Forests
3	3	Tropical Conifer forests
4	4	Temperate Broadleaf & Mixed Forests
5	5	Temperate Conifer forests
6	6	Boreal Forests
7	7	Tropical Grasslands & Savannas
8	8	Temperate Grasslands & Savannas

2. Use **Edit > Copy** to copy both the code column (MHT in this example) and the description column (Habitat Type) from the Summary table to the clipboard.
3. In the source (original) table, double-click on the code column to see its Column Info dialog. Select **Value Labels** in the Column Properties drop-down menu.
4. Click in the Value Labels box to highlight it.

Column Properties

Value Labels

optional item

Remove

Value Labels

If a column has value labels, and Use Value Labels is checked, the labels will be displayed wherever the column data are displayed.

optional item

Add

Change

Remove

Value

Label

☒ Use Value Labels

5. Use **Edit > Paste** to assign the set of labels.

Value Labels

If a column has value labels, and Use Value Labels is checked, the labels will be displayed wherever the column data are displayed.

1 = Tropical Moist Broadleaf Forests

2 = Tropical Dry Broadleaf Forests

3 = Tropical Conifer forests

4 = Temperate Broadleaf & Mixed Forests

5 = Temperate Conifer forests

6 = Boreal Forests

Add

Change

Remove

Value

Label

☒ Use Value Labels

6. Click **OK** and you're done

## Obtain and Interpret Odds Ratios for Interaction Terms

Duane Hayes, SAS Institute

Logistic regression models are widely used throughout industry and academia. They are appropriate when attempting to model a binary response such as Yes/No, Live/Die, or Good/Bad, or an ordinal response where ordering of the levels is important (Good/Better/Best or Mild/Moderate/Severe). Odds ratios in these models are used to interpret the effect of the factors included in the model. Odds ratios for interaction terms are more difficult to compute and to interpret because of the properties of the odds ratio. With an interaction term, there is no single odds ratio. To get an odds ratio for one of the main effects involved in the interaction, you must define a fixed level of the other effect. That is, the odds ratio for an effect A, when the interaction term A\*B is included in the model, is a function of a fixed level of B. Analyses in JMP are no different.

### Example with Interaction

To demonstrate, let's look at an example. The (simulated) data that is shown in *Figure 1* lists the **Gender** and **Age** of 100 subjects and whether the subject experienced low back **Pain** in the prior six months. The data is included on the web site with this issue of JMPer Cable at

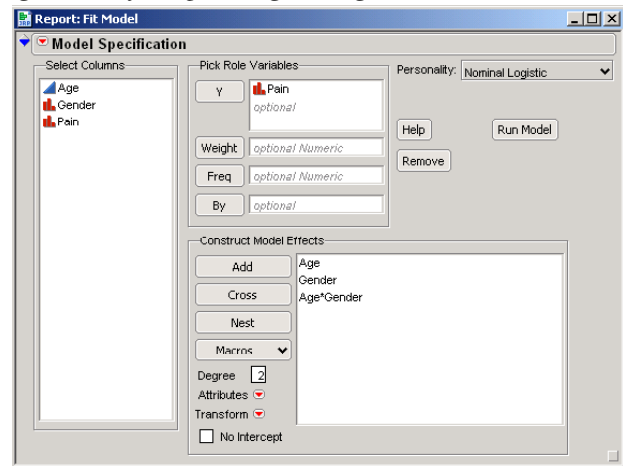
<http://www.jmp.com/about/newsletters/jmpercable/>

Figure 1 Partial Listing of Example Data

Odds Ratio Example				
Odds Ratio Example				
Model				
	Age	Gender	Pain	
	1	47 Male	No	
	2	69 Female	No	
	3	43 Male	Yes	
	4	61 Female	No	
	5	46 Male	No	
	6	63 Female	No	
	7	53 Male	Yes	
	8	42 Female	No	
	9	43 Male	Yes	
	10	68 Female	No	
	11	63 Male	Yes	
	12	51 Female	No	

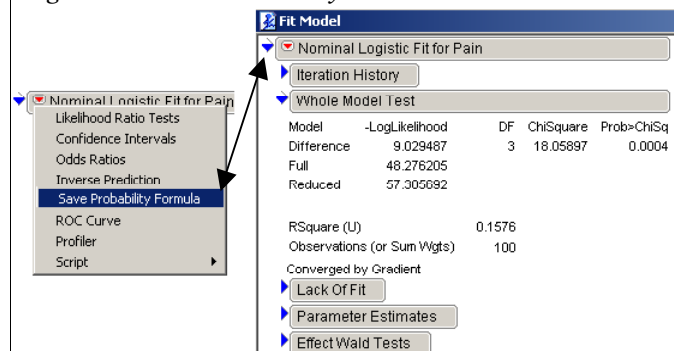
In this simple example, a logistic regression model can be fit with **Pain** as the response variable. The effects in the model are **Age**, **Gender**, and the **Age** by **Gender** interaction term. The interaction term is included because the belief exists that as age increases, males and females do not respond the same. The "Yes" response level for **Pain** is the event of interest. To analyze this model in JMP, choose **Analyze > Fit Model** and define the terms in the Fit Model dialog, as shown in *Figure 2*.

Figure 2 Defining the Logistic Regression Model



Click **Run Model** to see the results shown in *Figure 3*. Next, select **Save Probability Formula** from the pop-up menu on the Nominal Logistic title bar.

Figure 3. Save the Probability Formulas



This selection generates the following new columns in the original data table:

**Lin[Yes]** – the linear predictor, or the predicted logit function

**Prob[No]** – the predicted probability of the response being No

**Prob[Yes]** – the predicted probability of the response being Yes

**MostLikely Pain** – the predicted level of the response.

Each of these new columns has a formula. To obtain customized odds ratios, first create new rows in the data table and enter age and gender values for the comparison settings. For example, suppose you want to determine the odds ratio of 55-year-old males versus 65-year-old males, and 60-year-old males versus 60-year-old females. *Figure 4* shows the four additional rows with the formulas described above evaluated.

Figure 4 Add Rows for Values to Compute Customized Odds Ratios

	Age	Gender	Pain	Lin[Yes]	Prob[No]	Prob[Yes]	Most Likely Pain
93	47	Male	No	-0.7640531	0.68223306	0.31776694	No
94	59	Female	No	-2.4115511	0.9177039	0.0822961	No
95	44	Male	No	-0.940974	0.71929636	0.28070364	No
96	48	Female	No	-1.8124389	0.85965638	0.14034362	No
97	46	Male	No	-0.8230267	0.69487846	0.30512154	No
98	52	Female	No	-2.0302979	0.88394164	0.11605836	No
99	61	Male	Yes	0.06157756	0.48461047	0.51538953	Yes
100	55	Female	No	-2.1936921	0.89968163	0.10031837	No
101	65	Male		0.29747205	0.42617558	0.57382442	Yes
102	55	Male		-0.2922642	0.57255035	0.42744965	No
103	60	Male		0.00260394	0.49934901	0.50065099	Yes
104	60	Female		-2.4660159	0.9217248	0.0782752	No

The **Lin[yes]** column contains the formula for the linear combination of regressor terms needed to compute the probabilities of response values (No or Yes) for each observation. The odds ratio to compare two observations is the log of the ratio of their **Lin[Yes]** values. For example, the odds ratio for 65- versus 55-year-old males is computed:

$$\begin{aligned}\text{Odds Ratio} &= \text{Exp}(0.29747205 - (-0.2922642)) \\ &= \text{Exp}(0.58973625) = 1.8035\end{aligned}$$

This odds ratio of 1.8035 indicates that the odds of pain increase by a factor of 1.8035 for males as age increases from 55 to 65.

The second odds ratio of interest (60-year-old males versus 60-year-old females) is computed as:

$$\begin{aligned}\text{Odds Ratio} &= \text{Exp}(0.00260394 - (-2.4660159)) \\ &= \text{Exp}(2.46861984) = 11.8061\end{aligned}$$

This odds ratio of 11.8061 indicates that the odds of pain for 60-year-old males is 11.8061 times higher than for 60-year-old females.

These computations demonstrate a method of obtaining customized odds ratios for logistic regression models using JMP. This method is not limited to interaction

terms. It can be computed for specific levels of main effects. So, the default odds ratios given in JMP is just fine in some situations, but in some cases the tests of interest must be defined by hand. The method described here always gives the comparison of interest.

### Script for Specific Odds Ratio Computation

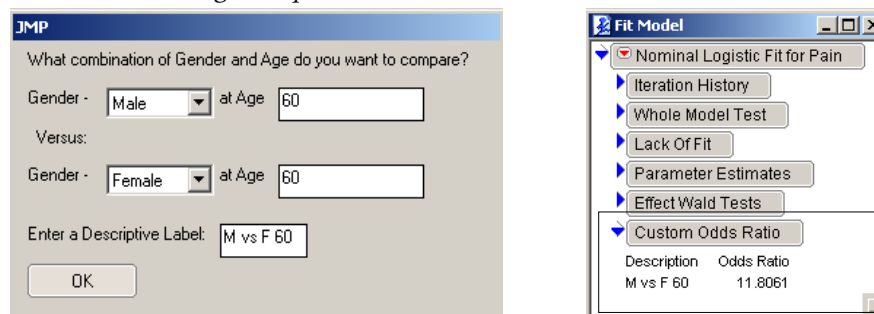
These computations can also be done with the JMP Scripting Language (JSL). A script called **Custom Odds Ratio.jsl** is available with the data at

<http://www.jmp.com/about/newsletters/jmpcable>

First open the **Odds Ratio Example.jmp** table, and then run the script. The script performs the nominal logistic regression and creates a dialog that prompts for the comparison you want to see. The custom odds ratio results appear at the bottom of the nominal logistic report.

To compare 60-year-old males to 60-year-old females, complete the dialog provided by the script, as shown on the left in *Figure 5*. When you click **OK**, the outline node called **Custom Odds Ratio** appears in the logistic regression report. The odds ratio comparing pain for these two groups is 11.8061, as calculated manually.

Figure 5 Find Custom Odds Ratio Using a Script



## Do You Know About Trainer's Kits?

Many companies are faced with bringing new JMP users up to speed. A broad curriculum of JMP training is available through SAS Institute as public courses scheduled regularly at SAS training centers worldwide or onsite at your facility.

However, if you are a company with instructors on your staff who already know JMP, you can choose any course from our current JMP curriculum and we will create a JMP Trainer's Kit for you. The Trainer's Kit includes:

- presentation materials
- thirty copies of the course notes
- course data
- a free seat in the corresponding public class so your instructor can see a JMP instructor present the material.

There are many benefits to using a professionally prepared JMP Trainer's Kit, such as:

- having the same high-quality training materials that JMP instructors use
- using your own training staff efficiently
- using materials written by the experts in JMP who have years of technical training experience.
- having complete ready-to-use materials
- being able to make modifications to the material so that they to suit your own needs.

The following courses are available for JMP Software as public courses, onsite courses, or Trainer's Kits.

ANOVA and Regression, for either version 5 or version 6 of JMP  
 Analysis of Attribute Data  
 Design and Analysis of Experiments, for either version 5 or version 6 of JMP  
 Identifying and Modeling Process Cycles  
 Introduction to the JMP Scripting Language  
 Modern Design of Experiments  
 New Features in JMP6  
 Reliability Analysis  
 Statistical Data Exploration, for either version 5 or version 6 of JMP  
 Statistical Quality control  
 JMP and Statistical Essentials for SAS Microarray Solution

For a list and descriptions of all the courses offered by SAS Institute, go to

<http://support.sas.com/training/us/heclist.html>

If you have questions, or would like to order a JMP Trainer's kit, contact Deborah Upchurch in SAS Education at (919) 531-7312 or send email to [training@jmp.com](mailto:training@jmp.com)

## SAS Training Facilities in North America

Public courses taught by SAS Institute instructors for SAS and JMP are held in training centers worldwide. For more information about courses, locations, and schedules in North America, see the map at

[http://support.sas.com/training/map\\_na.html](http://support.sas.com/training/map_na.html)

and click on a SAS training site for more information.



## Upcoming Conferences

**July 25, 2006**

Lean University—The Tools of Reliability  
Cleveland, OH

**August 7-10, 2006**

Drug Discovery Technology & Development  
World Congress  
Boston, MA

**September 7-10, 2006**

Designed Experiments: Recent Advances in  
Methods and Applications  
Southampton, UK

**September 17-20, 2006**

NESUG  
Philadelphia, PA

**September 25-28, 2006**

Discovery-2-Diagnostics Conference & Expo  
Boston, MA

**September 26-29, 2006**

Lean Six Sigma Summit West 2006  
Las Vegas, NV

**September 27-29, 2006**

WUSS  
Irvine, CA

**October 8-10, 2006**

SESUG  
Atlanta, GA

**October 9-13, 2006**

American Society of Human Genetics  
New Orleans, LA

**October 12-13, 2006**

50th Annual Fall Technical Conference  
Columbus, OH

**October 15-17, 2006**

SCSUG  
Irving, TX

**October 22-24, 2006**

MWSUG  
Dearborn, MI

**October 29-31, 2006**

PNWSUG  
Seaside, OR



There are a new set of flash tutorials available on the JMP website at

<http://www.jmp.com/software/demos.shtml>

These short non-interactive tutorials instruct you in sample methods for exploring data and show how to accomplish different tasks you may find valuable during data analysis. The tutorials are completely automated using Macromedia Flash Player 7 or higher. Simply click on the tutorial name and watch the automated demo.

**Task:** Create a table to show the mean and standard deviation for height and weight for male and female students.

**Tutorial:** Using Tabulate to produce summary statistics.

**Task:** Compare crash test results for cars made in 1990-1991. Which cars have the worst results for injuries to the left leg and to the head?

**Tutorial:** Use Tree Map to compare relationships between variables.

**Task:** Quit JMP and automatically retrieve all the data tables and report windows you had open when you return later.

**Tutorial:** Saving JMP work sessions.

**Task:** Integrate a PowerPoint presentation into JMP so you can launch JMP to support your presentation points without switching between applications.

**Tutorial:** PowerPoint to JMP.

**Task:** Develop a single presentation that involves multiple JMP data tables, statistical reports, pictures, and HTML pages, taking advantage of the live connection between tables and reports.

**Tutorial:** Presentations in JMP.

**Task:** Organize many different reports to run for each project so that you don't need to remember the command sequence every time you run the reports.

**Tutorial:** Create a set of customized menus.



## Spline Models

Lee Creighton, SAS Institute

In a review paper on spline models, Smith (1979) gives the following definition:

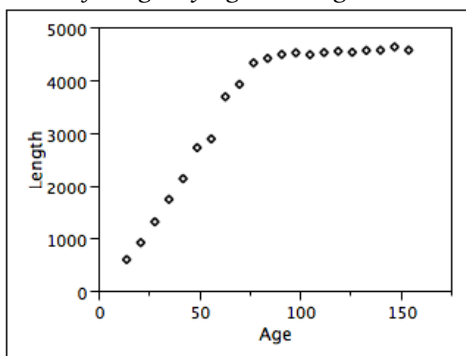
“Splines are generally defined to be piecewise polynomials of degree  $n$  whose function values and first  $(n-1)$  derivatives agree at points where they join. The abscissas of these joint points are called knots. Polynomials may be considered a special case of splines with no knots, and piecewise (sometimes also called grafted or segmented) polynomials with fewer than the maximum number of continuity restrictions may also be considered splines. The number and degrees of polynomial pieces and the number and position of knots may vary in different situations.”

First consider splines with known knots, that is, splines for which the values of the independent variable are known for the joint points. Fitting spline models with known knots is much easier than fitting them with unknown knots, because with known knots you can use linear regression methods. Estimation of spline models with unknown knots requires the use of nonlinear methods.

As an example, consider the data tables called **Fish.jmp**, with variables **Age** and **Length**, which can be found at <http://www.jmp.com/about/newsletters/jmpercable>

Judging from the plot in *Figure 1*, the rate of increase in **Length** is roughly constant until about **Age** = 80. The graph clearly shows a knot point when age is about 80, at which point growth appears to stop abruptly. A linear spline with knot at **Age** = 80 would be suitable in this situation.

Figure 1 Plot of Length by Age Showing Knot Point



In order to perform a regression analysis, you need a linear regression equation to represent the spline model. For this equation, define a new variable in the data table (call it **AgePlus**) as the maximum of **Age** – 80 and zero.

- Double-click to the right of the last column in the table to add a new column.
- Name the new column **AgePlus**.
- Right-click (Control-click on the Macintosh) on the new column and select **Formula** from the menu that appears.
- In the Formula Editor, choose the **Maximum** function from the **Statistical** functions and enter:

Maximum(Age - 80, 0)


Note: The **Maximum** function initially shows only a single argument. To add a second argument, press the comma key or use the insert button (  ).

Figure 2 shows the **Fish** data table with computed values for the new variable **AgePlus**.

Figure 2 Partial Listing of Data Table with New Variable

	Age	Length	AgePlus
6	49	2725	0
7	56	2890	0
8	63	3685	0
9	70	3920	0
10	77	4325	0
11	84	4410	4
12	91	4485	11
13	98	4515	18
14	105	4480	25
15	112	4520	32

The spline model is represented with the regression equation

$$\text{Length} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{AgePlus}) + \varepsilon$$

For **Age** < 80, that is

$$\text{Length} = \beta_0 + \beta_1(\text{Age}) + \varepsilon$$

For **Age** > 80, the equation can be written

$$\text{Length} = (\beta_0 - 80\beta_2) + (\beta_1 + \beta_2)(\text{Age}) + \varepsilon$$

Notice that both expressions give the same estimated **Length** when **Age** = 80. In other words the two line segments are joined at **Age** = 80. Now you can do the regression analysis with the Fit Model platform using both **Age** and **AgePlus** as effects:

- Choose **Analyze > Fit Model**.
- Assign **Length** to the Y role.

- Assign **Age** and **AgePlus** as effects.
- Click **Run Model** to see the results in *Figure 3*.

The plot of actual values against predicted values shows a good fit of the model to the data. The fitted equations are:

$$\text{Length} = -327.62 + 60.18(\text{Age}) - 58.86(\text{AgePlus})$$

Therefore, the fitted spline model for  $\text{Age} < 80$  is

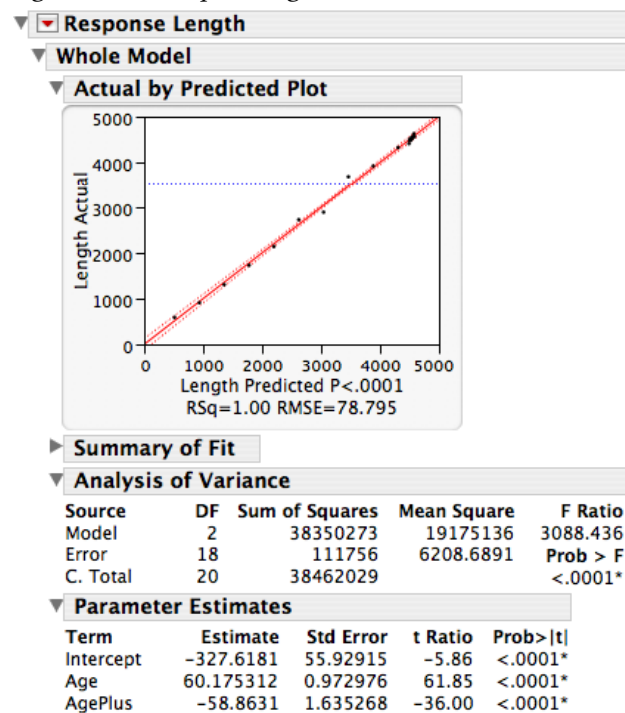
$$\text{Length} = -327.62 + 60.18(\text{Age})$$

For  $\text{Age} < 80$  the model is

$$\begin{aligned} \text{Length} &= (-327.62 + 58.86 \cdot 80) + (60.18 - 58.85)(\text{Age}) \\ &= 4376.38 + 1.32(\text{Age}) \end{aligned}$$

Notice the small slope of 1.32 for  $\text{Age} > 80$ .

*Figure 3 Linear Spline Regression*



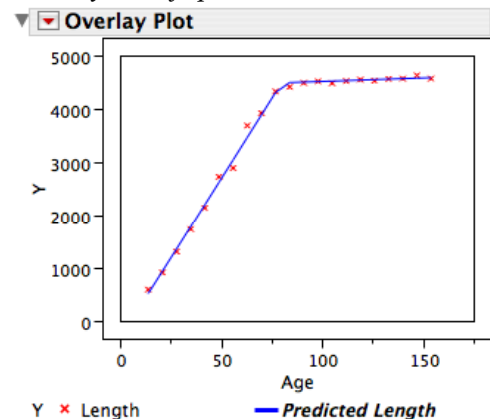
To see a plot of the fitted values to the actual data, save the predicted values and use the Overlay Plot platform.

- Choose **Save Columns > Predicted Values** from the Fit Model platform menu to create the new column, **Predicted Length**, in the data table.
- Choose **Graph > Overlay Plot**
- Assign **Length** and **Predicted Length** as Y.
- Assign **Age** as X.
- Click **OK** to see the overlay plot in *Figure 4*.

To see the options in effect in *Figure 4*,

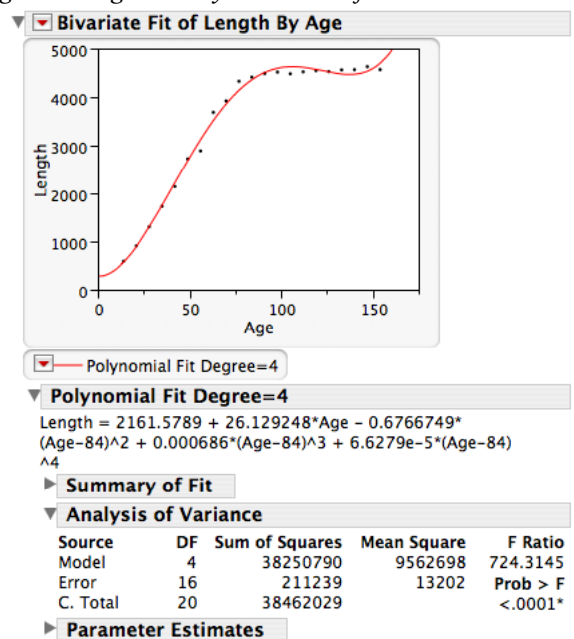
- Right-click (Control-click on the Macintosh) on the **Predicted Length** legend at the bottom of the plot.
- Select **Connect Points** from the menu that appears.
- Repeat the same step to uncheck **Show Points**.

*Figure 4 Overlay Plot of Spline Fit*



A formal test is not available but you can compare the fit of the linear spline to a fit of a polynomial. *Figure 5* shows a 4<sup>th</sup> degree polynomial fit of the data, using the Fit Y by X platform. In *Figure 3* you see Mean Square (Error) = 6,208. In *Figure 5* the Mean Square (Error) = 13,202 for the 4<sup>th</sup> degree polynomial, indicating a much better fit for the spline function.

*Figure 5 Degree 4 Polynomial Fit of Data*





JMP  
SAS Campus Drive  
Cary, NC 27513 USA  
Tel: (919) 677-8000

**About JMPer Cable**

Issue 20 Summer 2006

JMPer Cable is mailed to JMP users who are registered users with SAS Institute. It is also available online at [www.jmp.com](http://www.jmp.com)

**Contributors**

Mark Bailey, Lee Creighton,  
Duane Hayes, Diana Levey

**Editor**

Ann Lehman

**Printing**

SAS Institute Print Center

**Questions, comments, or for  
more information about JMP, call**

1-877-594-6567

or visit us online at

[www.jmp.com](http://www.jmp.com)

**To Order JMP Software**

1-877-594-6567

Copyright© 2006 SAS Institute Inc. All rights reserved. SAS, JMP, JMPer Cable, 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. Six Sigma is a registered trademark of Motorola, Inc.