



JMP® Pro 13 Modeling Workflow Enhancements With Predictor Screening, Generalized Regression and the New Formula Depot

Karen Copeland – Boulder Statistics

Abstract

- Building models is often an exploratory, iterative process that can result in many saved models. The new Formula Depot in JMP13 Pro provides a workflow to save, compare, and score models without cluttering your data table.
- Prior to building models one may need to screen a large number of factors to identify those with modeling potential. The JMP predictor screening platform is one tool for the job.
- Once a smaller set of factors is identified then the generalized regression personality in the fit model platform is a tool for model building and further factor selection. Generalized regression options include various algorithms and validation options to facilitate factor selection and/or minimize over fitting.
- The generalized regression capabilities have been expanded in JMP Pro 13 to include more types of models, more fitting algorithms and model visualizations, such as ROC curves and profilers.

Predictor Screening

- Predictor screening uses bootstrap forests to identify strong predictors from a pool of predictors.
- Because of the randomness in the procedure, results from re-running the analysis will shift. The larger the set of predictors, the more variability you will see in the results across runs.
- I typically run predictor screening at least 3 times and look at the overall list.

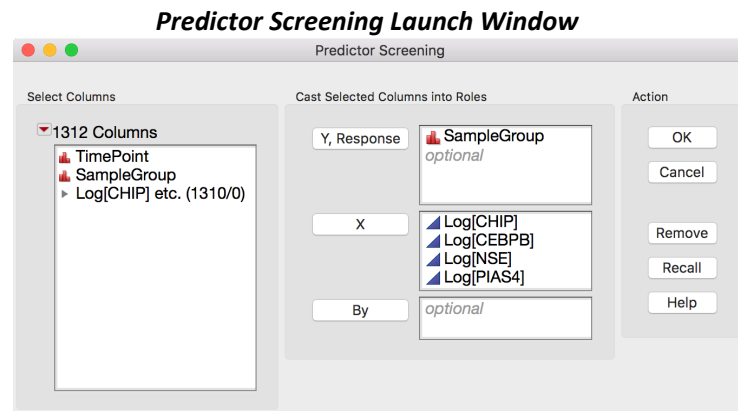
Generalized Regression

- The generalized regression platform is my go-to platform for model building.
- Workflow improvements in JMP13 include
 - Re-launch with active effects
 - ROC curves (for categorical responses)

Formula Depot

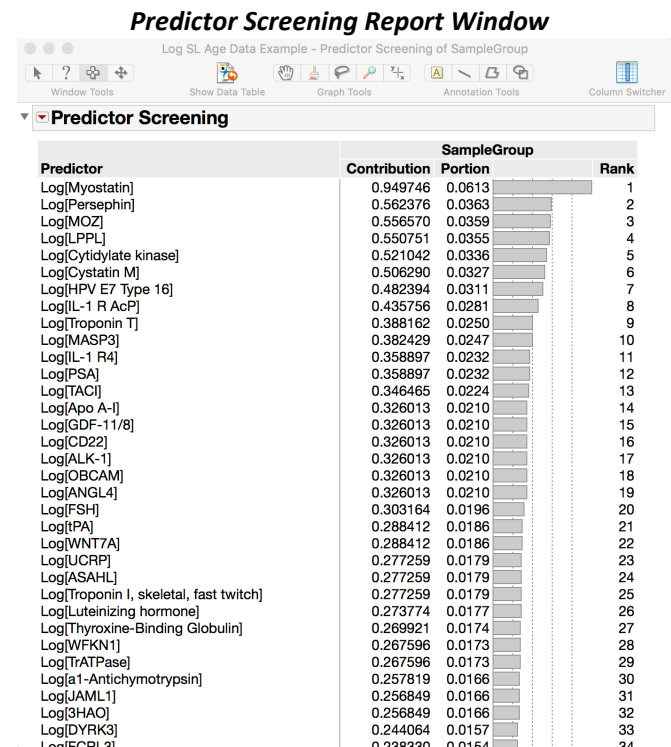
- The FD is a “container” (a *.jrp file) that holds formulas (remember a model is a formula).
- From many modeling platforms you now have the option to “publish” your model. This saves the prediction formula to the Formula Depot rather than to your data table.
- Once you have a models in the FD you can compare them with the model compare or profiler.
- You can run the model to add it to your data table or to add it to any data table. However, be careful, if your column names or data structures change between tables your model will not run. For example, if you build a model with a gender column coded as M and F but try to run it in a data table with gender as Male and Female the model will not run. Recode is a simple fix.
- For those who need a model outside of JMP you can score the model from the FD. This will translate the JSL code into another coding language (C, Java, Python, SQL, or SAS) for implementation outside of JMP.

Analyze > Screening > Predictor Screening



Y, Repsones = SampleGroup = Gender (10 male, 10 female)
X = Predictors = 1310 proteins

The report window lists the predictors based on their contribution from a bootstrap forest model. I use this as an exploratory tool to help select predictors. I am most interested in the top predictors. The Portion column and corresponding bar chart help to identify how many of the predictors I may wish to explore further.



Analyze > Fit Model : Generalized Regression Personality

The image shows two screenshots from a software interface. The left screenshot, titled "Model Launch Window", shows the "Fit Model" dialog. The "Personality" is set to "Generalized Regression", "Distribution" is "Binomial", and "Target Level" is "F". The "SampleGroup" variable is selected as the response. The "Construct Model Effects" section shows a list of predictors: Log[Myostatin], Log[Persephin], Log[Apo A-I], Log[GDF-11/8], Log[FSH], Log[LKHA4], Log[TrATPase], Log[LRP8], Log[CHL1], and Log[MASP3]. The right screenshot shows the "Generalized Regression for SampleGroup = F" model launch window. It displays the "Singularity Details" section with "Lasso" as the Estimation Method and "Adaptive" checked. The "Validation Method" is set to "AICc". A blue arrow points from the "Personality" dropdown in the left window to the "Generalized Regression" title in the right window.

Model Launch Window

Select Columns: 2628 Columns

- PlatePosition
- SampleID
- Barcode
- TimePoint
- ExtIdentifier
- SampleGroup
- Subject_ID
- SampleUniqueID
- CHIP etc. (1310/0)
- Log[CHIP] etc. (1310/0)

Pick Role Variables

Y: SampleGroup

Weight: optional numeric

Freq: optional numeric

Validation: optional

By: optional

Construct Model Effects

Add: Log[Myostatin], Log[Persephin], Log[Apo A-I], Log[GDF-11/8], Log[FSH], Log[LKHA4], Log[TrATPase], Log[LRP8], Log[CHL1], Log[MASP3]

Cross:

Nest:

Macros:

Degree: 2

Attributes:

Transform:

No Intercept:

Personality: Generalized Regression

Distribution: Binomial

Target Level: F

Help, Run, Recall, Keep dialog open, Remove

Generalized Regression for SampleGroup = F

Model Launch

Singularity Details

Estimation Method: Lasso

Adaptive: ☒

Advanced Controls

Validation Method: AICc

Early Stopping: ☐

Go

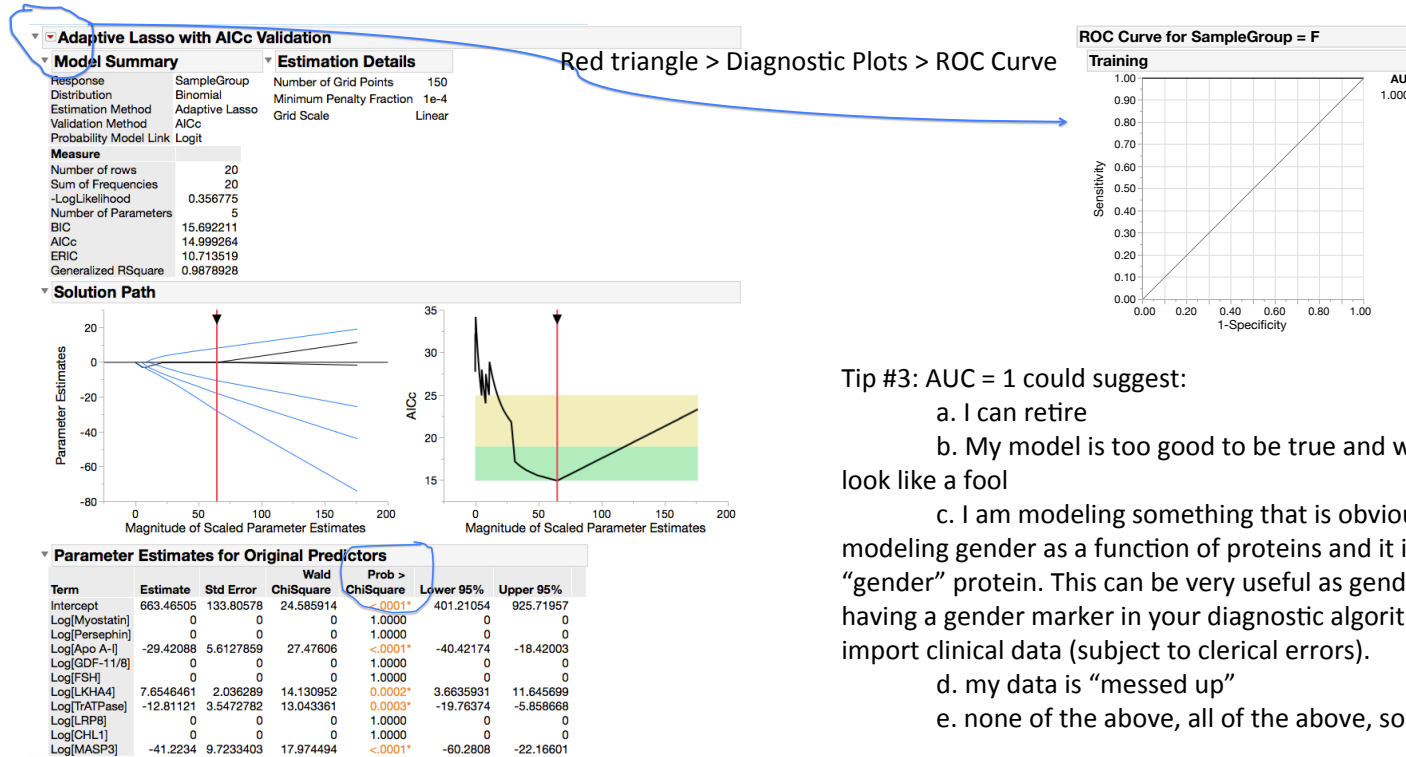
Next Slide

Tip #1: Predictor Screening report is linked to the data table, highlight column in the report to select in the data table then click Add in the model launch window.

Tip #2: Save the predictor screening results to a data table (right click > Make into Data Table) then copy column names from this table and paste into the model dialog.

Tip #3: Select columns of interest and then use column reorder to place them together in your data table.

Analyze > Fit Model : Generalized Regression Personality



Tip #3: AUC = 1 could suggest:

- I can retire
- My model is too good to be true and when I score verification data I will look like a fool
- I am modeling something that is obvious such as here where I am modeling gender as a function of proteins and it is not surprising that we can find a "gender" protein. This can be very useful as gender could impact your outcome and having a gender marker in your diagnostic algorithm keeps you from having to import clinical data (subject to clerical errors).
- my data is "messed up"
- none of the above, all of the above, some mix of the above

Tip #1: Click on column heading to sort by that column.

Tip #2: Red triangle > Regression Reports > Active Parameter Estimates.

Tip #4: Red triangle > Relaunch with Active Effects.

Analyze > Predictive Modeling > Formula Depot

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-44.88304	3.8514794	135.48464	<.0001*	-523.7916	-97.2164
Log[PIANP]	-15.87713	3.9522338	37.19159	<.0001*	12.085204	-37.67812
Log[PARC]	14.453661	2.4618091	34.470452	<.0001*	9.6286038	19.278718
Log[MIC-1]	29.227663	4.3167483	45.84318	<.0001*	20.766992	37.688334

Red triangle > Save Columns > Publish Prediction Formula

Report: Formula Depot

Window Tools Show Data Table

Formula Depot


Manage Models and Generate Scoring Code

Formula Scripts

Fit Generalized - SampleGroup

Formulas

Table	SL Age Data Example.jmp
Creator	Fit Generalized
Target	SampleGroup
Created	Most Likely SampleGroup, Probability(SampleGroup=F), Probability(SampleGroup=M)
Factors	Log[Apo A-I], Log[LKHA4], Log[MASP3], Log[TrATPase]
Others	



The screenshot shows the 'Add Formula from Column' menu in SAS Studio. The menu is open, displaying a list of options. A blue circle is drawn around the 'Generate C Code' option, which is the first option in the highlighted section. Other options include 'Show Scripts', 'Copy Scripts', 'Copy Formulas as Transforms', 'Run Scripts', 'Generate Python Code', 'Generate JavaScript Code', 'Generate SAS Code', 'Generate SQL Code', 'Model Comparison', 'Profiler', 'Redo', and 'Save Script'.

- Add Formula from Column
- Show Scripts
- Copy Scripts
- Copy Formulas as Transforms
- Run Scripts
- Generate C Code
- Generate Python Code
- Generate JavaScript Code
- Generate SAS Code
- Generate SQL Code
- Model Comparison
- Profiler
- Redo
- Save Script

- Code generation for use of models outside of JMP!

Red triangle > Run Script

- Show Script
- Copy Script
- Copy Formula
- Copy Formula as Transform
- Rename New Column**
- Generate C Code
- Generate Python Code
- Generate JavaScript Code
- Generate SAS Code
- Generate SQL Code
- Run Script
- Remove

Red triangle > Rename New Column

Please Enter Values

Rename formula column scripts

Probability(SampleGroup=F)	Model 1 Pr(F)
Probability(SampleGroup=M)	Model 1 Pr(M)
Most Likely SampleGroup	Model 1 ML

Cancel OK

Model 1 Pr(F)	Model 1 Pr(M)	Model 1 ML
4.0534e-24	1 M	
0.99982806	0.00017194 F	
2.314e-20	1 M	
3.1773e-20		
0.00000000	... 04e-15 F	
0.9999489	5.10535e-6 F	

Adds model to any JMP table. The table needs the model input columns.

Red triangle > Show Scripts

```

1 // Fit Generalized Formulas
2 New Column("Probability1 SampleGroup#") =
3     Numeric
4     Formula
5
6     1768.56882685363 - .75.344937124313 * %Name("Log[Apo A-1]")
7     + 15.018536881947 * %Name("Log1[UKM41]") + -21.3359593846483
8     * %Name("Log1[ITrAse1]") + -112.749755058607 * %Name("Log[MAS3])
9
10
11     }
12     Set Property(
13         {"Response Probability",
14          {"SampleGroup", "Creator", "Fit Generalized"}
15         Model Name "Adaptive Lasso", ID1 1227813665 }
16     )
17 New Column("Probability1 SampleGroup#") =
18     Numeric
19     Formula
20
21     - Log(1768.56882685363 - .75.344937124313 * %Name("Log[Apo A-1]")
22     + 15.018536881947 * %Name("Log1[UKM41]") + -21.3359593846483
23     * %Name("Log1[ITrAse1]") + -112.749755058607 * %Name("Log[MAS3])
24
25
26     }
27     Set Property(
28         {"Response Probability",
29          {"SampleGroup", "M", "Creator", "Fit Generalized"}
30         Model Name "Adaptive Lasso", ID1 1227813665 }
31     )
32 New Column("Most Likely SampleGroup",
33     Character
34     Nominal
35     Formula
36
37     IF(%Name("Probability1 SampleGroup#") =
38     %Name("Probability1 SampleGroup#M") =
39     "M",
40
41
42
43

```