

Using Generalized Regression to Analyze Designed Experiments

11/5/2021 Developer Tutorial

Clay Barker

Why do we design experiments?

From asq.org/quality-resources

It allows for multiple input factors to be manipulated, determining their effect on a desired output (response). By manipulating multiple inputs at the same time, DOE can identify important interactions that may be missed when experimenting with one factor at a time.

A well-performed experiment may provide answers to questions such as:

- 1. What are the key factors in a process?*
- 2. At what settings would the process deliver acceptable performance?*
- 3. What are the key, main, and interaction effects in the process?*

Why do we design experiments

Let's focus on that second part...

1. *What are the key factors in a process?*
2. *At what settings would the process deliver acceptable performance?*
3. *What are the key, main, and interaction effects in the process?*

#1 and #3 sound a lot like explaining.

#2 sounds a lot like predicting.

The generalized regression platform in JMP Pro makes it easy to tackle all 3.

Why do we design experiments?

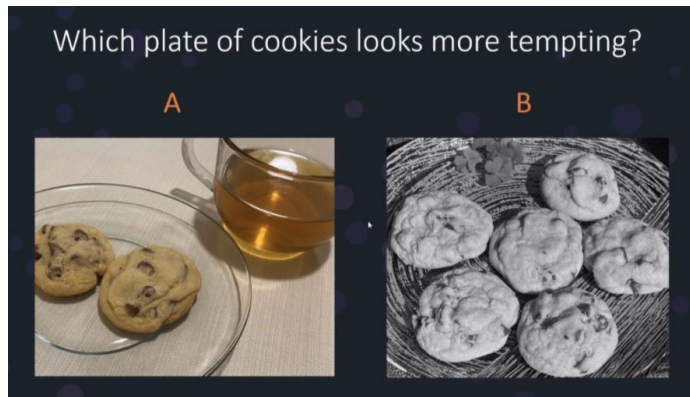
An example

The Great Cookie Experiment at Discovery Europe 2021

Factors

Add Factor ▼ Remove Add N Factors 1

Name	Role	Changes	Values		
▼ Beverage:	Categorical	Easy	Milk	Coffee	Tea
▼ Type of Plate:	Categorical	Easy	Solid		Pattern
▼ Setting:	Categorical	Easy	Inside		Outside
▼ Number of Cookies on Plate:	Categorical	Easy	2		6
▼ Color:	Categorical	Easy	Black & White		Color
▼ Use a Prop:	Categorical	Easy	Yes		No



What is Generalized Regression?

Generalized Regression is a personality of Fit Model available in JMP Pro. We call it Genreg for short.

What is it?

Genreg is the place to build Generalized Linear Models in JMP Pro.

Key capabilities

1. Generalized Linear Models
2. Variable Selection
3. Interactivity (of course 😊)

What is Generalized Regression

The methods available in Genreg are good for both analyzing both observational data and results from a designed experiment.

The variable selection methods in Genreg help us

- Identify the most important factors that impact our response.
- Make better predictions.

...these qualities line up pretty perfectly with our goals when designing an experiment.

What is Variable Selection?

Variable selection is the process of selecting a subset of variables (predictors) to use in modeling a response variable.



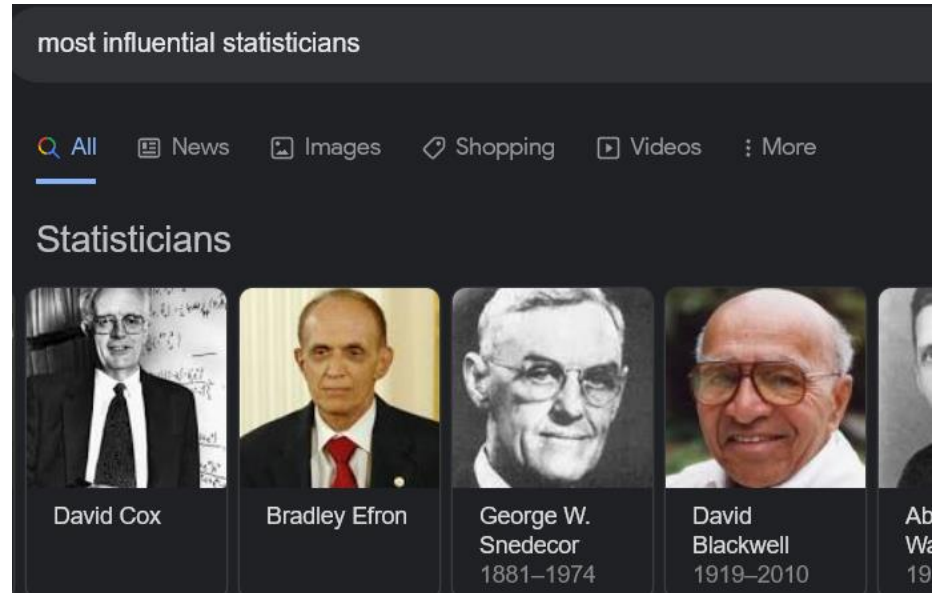
- We have a candidate set of explanatory variables that may be associated with the response. Throw them all into a variable selection procedure and see what happens.
- But automation doesn't give us a free pass to not think about our data!

What is Variable Selection?

It's kind of a big deal

Variable selection is a very active area of research.

In an interview in the late 90s, Brad Efron of Stanford said that variable selection in regression is the most important problem in statistics.



What is Variable Selection?

There are tons of variable selection methods out there...

Many methods follow this general strategy:

1. Fit a sequence of models (maybe a **lot** of models)
2. Use some metric to see how well each model fits our data.
3. Keep the model that fits best based on (2).

Outline

Today we'll talk about how the variable selection methods in Genreg apply to designed experiments.

1. How do we compare different models?
2. How do we decide on the sequence of models to fit?
3. Some example problems



How do we compare models?

Comparing Models

Cross-validation

How well does a model fit?

One way to think of this: how well would our model fit new observations?

We can approximate this by breaking our data into parts.

Use one part to estimate our model.

Apply that model to the other part.

This is called *cross-validation*.

	X1	X2	X3	X4	X5	X6	X7
1	1	0	1	1	-1	1	-1
2	-1	-1	1	1	0	-1	-1
3	1	-1	0	1	1	-1	1
4	1	1	-1	1	-1	-1	0
5	1	1	-1	-1	0	1	1
6	0	-1	-1	-1	-1	-1	-1
7	-1	-1	-1	1	-1	1	1
8	0	0	0	0	0	0	0
9	-1	1	1	0	-1	-1	1
10	-1	1	-1	1	1	0	-1
11	-1	0	-1	-1	1	-1	1
12	-1	-1	1	-1	1	1	0
13	1	-1	-1	0	1	1	-1
14	-1	1	0	-1	-1	1	-1
15	1	1	1	-1	1	-1	-1
16	0	1	1	1	1	1	1
17	1	-1	1	-1	-1	0	1

Training

Validation

Comparing Models

Cross-validation

But wait...Should we break up our designed experiment?

We worked pretty hard to get that structure.

Dropping just one observation can be bad.

	X1	X2	X3	Y
1	0.16	0	-1	20.8053908
2	-1	1	-1	23.260038
3	-1	1	1	24.0807666
4	-0.54	-1	1	10.7704387
5	-1	0	0	15.4734538
6	-1	-1	-1	11.1681859
7	1	-1	0	18.0764967
8	0	1	0	26.2513821
9	1	1	-1	28.5406999
10	1	1	1	28.3694953

Response Y				
Singularity Details				
Term	Details			
Intercept	= - 0.27669*X1 + 0.08496*X2 + 0.1...			
Effect Summary				
Parameter Estimates				
Term		Estimate	Std Error	t F
Intercept	Biased	-1.509606	.	.
X1	Biased	0.2586707	.	.
X2	Biased	0.9902136	.	.
X3	Biased	1.4411947	.	.
X1*X2	Biased	1.4881684	.	.
X1*X3		0.980407	.	.
X2*X3	Biased	0.0129198	.	.
X1*X1	Biased	2.9374886	.	.
X2*X2	Biased	2.6328984	.	.
X3*X3	Zeroed	0	.	.

Comparing Models

Information Criteria

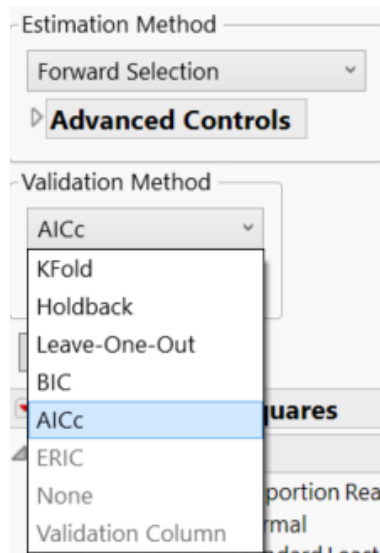
We need to be very careful using CV with designed experiments.

Instead, we suggest using an information criteria like the AIC or BIC.

They operate on the full data.

Cross-validation

Information criteria



Comparing Models

AICc and BIC

We interpret the AICc and BIC as a distance from the “true model”.
...so smaller is better when comparing them.

$$\text{AICc} = 2p - 2 * \log(L) + \frac{2p^2 + 2p}{n - p - 1}$$

$$\text{BIC} = p * \log(n) - 2 * \log(L)$$

n = number of observations

p = number of parameters

L = likelihood of fitted model

The likelihood function measures how well our models fit.

So the AICc and BIC are a compromise between how well a model fits and the complexity of our model.

Comparing Models

AICc and BIC in practice

After fitting a handful of models, we can use the AICc or BIC to choose a best model of the group.

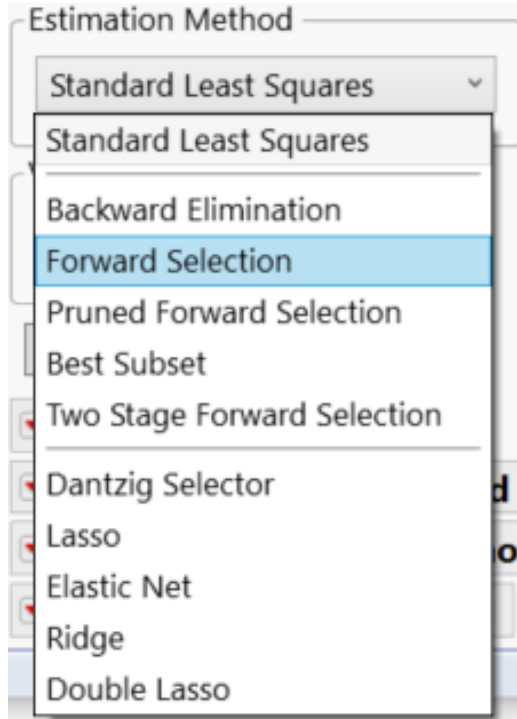
Model Comparison					
Show	Response Distribution	Estimation Method	Nonzero Parameters	AICc [^]	BIC
<input checked="" type="checkbox"/>	Normal	Standard Least Squares	17	-73.91236	-92.70914
<input checked="" type="checkbox"/>	t(5)	Maximum Likelihood	17	-73.01271	-91.80949
<input checked="" type="checkbox"/>	Gamma	Maximum Likelihood	17	-68.28429	-87.08107
<input checked="" type="checkbox"/>	Beta	Maximum Likelihood	17	-49.94901	-68.74579

We tend to recommend using the AICc with designed experiments because it includes a term that corrects for small sample sizes.



How do we do variable selection in Genreg?

Variable Selection in Genreg



These are Stepwise methods.
We tend to recommend these for
designed experiments.
Easy to explain and interpret.

These are penalized regression methods.
Great, but more for observational data.

We'll cover these next week.

Variable Selection in Genreg

What does a Selection Procedure Do?



A common strategy...

1. Start with nothing in the model (just an intercept)
2. Fit a sequence of models (usually of increasing size)
3. Stop when every effect is in the model

Then pick the “best” model based on a measure like AIC or cross-validation.

We could also do this in reverse (Everything -> ...drop stuff... -> Nothing)

Stepwise Methods in Genreg

Best Subset

Best subset (or all subsets) is exactly what it sounds like:

Given our predictors, fit every single model possible.

Keep the one that fits best.

The sequence of models we fit is all of the models.

Simple case with three candidate effects (X1, X2, X3):

Model Size	Models
0	Just the Intercept
1	(X1), (X2), (X3)
2	(X1, X2), (X1, X3), (X2, X3)
3	(X1, X2, X3)

We fit all 8 of these models and declare a best model based on AIC or BIC.

We win, right?

Stepwise Methods in Genreg

Best Subset

Awesome - we literally tried everything and we have the best.

Now consider a bigger model.

We have 10 continuous predictors.

Also want to consider two factor interactions and quadratic terms (RSM)

Main Effects + Interactions + Quadratics = 65 terms to consider in our model.

Now how many models do we need to fit?

$$2^{65} \approx 36900000000000000000$$

About 37 Quintillion!

Stepwise Methods in Genreg

Forward Selection

- Best Subset gets out of control fast
- Forward Selection uses heuristics to choose a good model of each size.
- Given our current model, what's the best effect to add to the model?

Simple and intuitive algorithm:

1. Start with just an intercept
2. Test each variable for inclusion. Add the variable with the smallest p-value.
3. Repeat (2) until everything enters or the model is saturated.

For k candidate effects, we end up with a sequence of $\min(n, k)$ fits;
Keep the best model based on AIC/BIC/CV.

Stepwise Methods in Genreg

FS Example

Consider a (very) simple example with 4 factors.

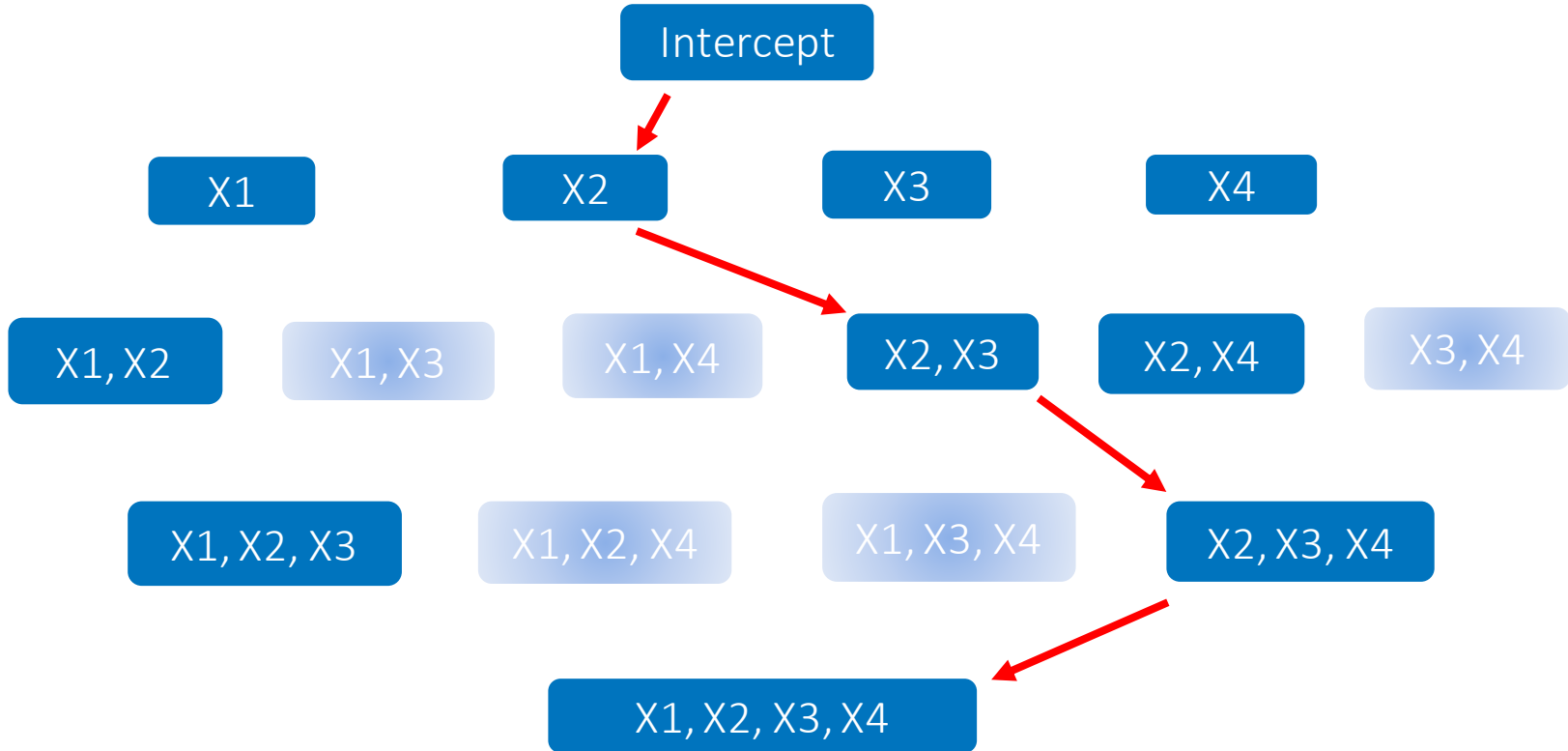
	Step 1	Step 2	Step 3	Step 4
<i>p</i> for X1	.2	.15	.2	.1*
<i>p</i> for X2	.001*			
<i>p</i> for X3	.6	.03*		
<i>p</i> for X4	.05	.3	.06*	

Now we have 5 models to fit and consider (out of 16 possible models)

1. Intercept only
2. X2
3. X2, X3
4. X2, X3, X4
5. X2, X3, X4, X1

Stepwise Methods in Genreg

Forward Selection



Stepwise Methods in Genreg

Backward Elimination

Pretty much FS in reverse

Fit a model, drop variables that aren't significant, refit, ...

BE algorithm

1. Start with everything in the model
2. Drop the worst effect based on p-values (bigger is worse)
3. Repeat (2) until we only have an intercept
4. Keep the best model in the sequence.

Not well defined when we can't fit the full model ($n < p$)

What does a large p-value really mean? Not much.

Stepwise Methods in Genreg

A Backwards Example

Another toy example with 4 predictors.

	Step 1	Step 2	Step 3	Step 4
p for X1	.001	.4*		
p for X2	.2	.15	.06*	
p for X3	.6*			
p for X4	.05	.03	.01	.003

Again, 5 models to fit and consider:

1. X1, X2, X3, X4
2. X1, X2, X4
3. X2, X4
4. X4
5. Just an intercept

Stepwise Methods in Genreg

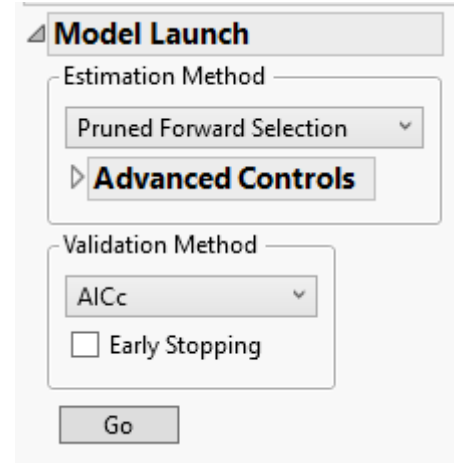
Forward and Backward Steps

There are good things about both Forward and Backward Selection.

Would it make sense to combine them?

That's our goal with the Pruned Forward Selection method in Genreg.

- Unique to Genreg (but similar to other methods)
- At each step in the algorithm, consider adding a term, dropping a term, or swapping terms.



The screenshot shows the 'Model Launch' dialog box in Genreg. It features a title bar with a collapse icon and the text 'Model Launch'. Below the title bar, there are two main sections: 'Estimation Method' and 'Validation Method'. The 'Estimation Method' section contains a dropdown menu currently set to 'Pruned Forward Selection' and a button labeled 'Advanced Controls' with a right-pointing triangle icon. The 'Validation Method' section contains a dropdown menu currently set to 'AICc' and a checkbox labeled 'Early Stopping' which is currently unchecked. At the bottom of the dialog box is a 'Go' button.

Stepwise Methods in Genreg

Pruned Forward Selection

Similar to what is often called Mixed Step selection other places.

The Algorithm starts similar to Forward Selection, but at each step

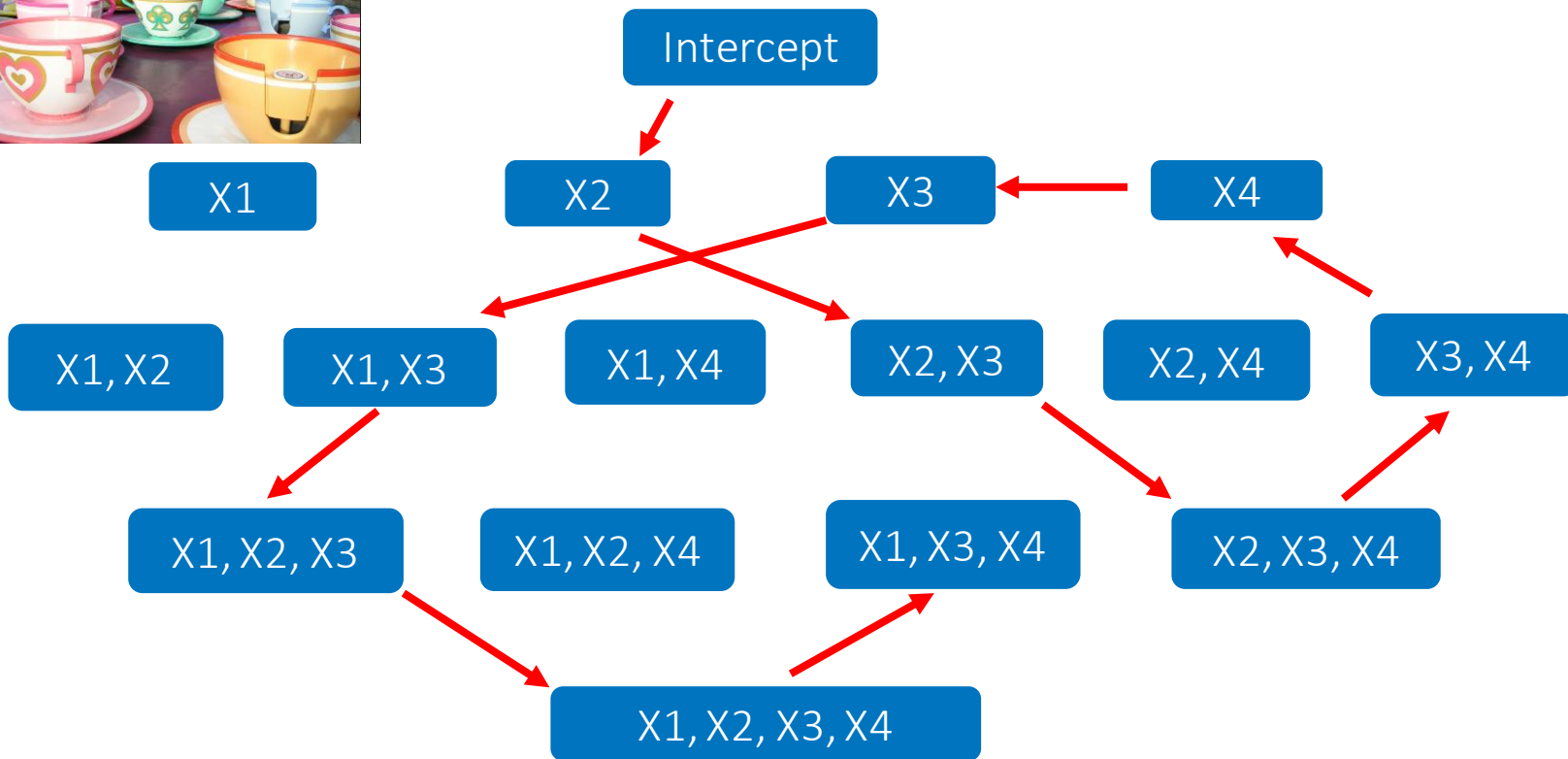
1. Find the variable that most wants to enter X_E (Score test)
2. Find the variable that most wants to leave X_L (Wald test)
 - A. Try adding X_E
 - B. Try removing X_L
 - C. Try swapping X_L for X_E
3. Go with A, B, or C based on which fits best.
4. Go back to 1.

Starts like FS but we can drop effects as we go.



Stepwise Methods in Genreg

PFS bounces around a lot

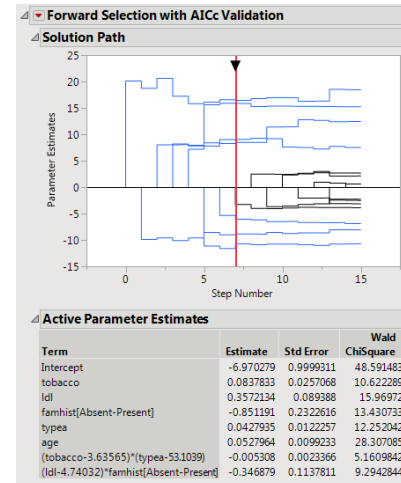
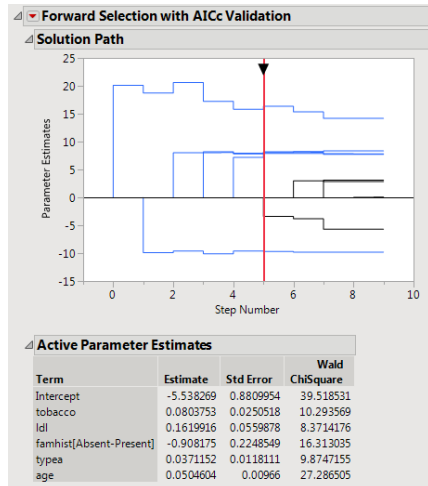


Stepwise Methods in Genreg

Two-Stage Forward Selection

We often fit models with main effects, interactions, and polynomials. It may make sense to break selection into two pieces:

1. Forward Selection just on main effects gives us an active set S .
2. Forward Selection on S and the higher order effects that contain S .

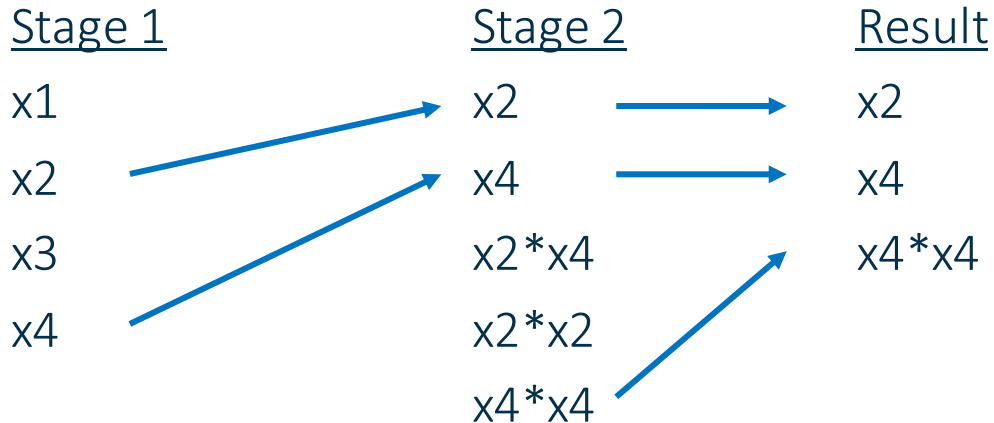


Stepwise Methods in Genreg

Two-Stage FS

Suppose we have 4 effects x_1 - x_4 .

Want to consider main effects, interactions, and squares.



This two stage process is especially nice if main effects and interactions are orthogonal (DSDs).

Selection methods in Genreg

...and one penalized regression method

- Candès and Tao (2007) suggested a new penalized regression method aimed at variable selection in the $n \ll p$ setting.

$$\hat{\beta}_{DS} = \arg \min_{\beta} \sum_j |\beta_j| \quad \text{subject to } |X^T(y - X\beta)|_{\infty} \leq s$$

- In words – control the magnitude of coefficients subject to a constraint on the maximum correlation between the design and the residuals.
- Kinda strange, but has shown promise for supersaturated designs.



DOE-specific options in Genreg

Which options are important for designs?

It's probably easiest just to look at examples...

Summary

The Generalized Regression platform in JMP Pro is the place to go when you need to analyze results of a designed experiment.

Some things to keep in mind...

- Probably stick with a stepwise method using AICc for validation.
- Probably want to use Effect Heredity (it will be checked by default).
- You may want a bigger model than the best model (especially in screening situations).



Thanks!
Clay.Barker@sas.com

sas.com