



# The Most Flexible Modeling Platform That You're Not Using (...but hopefully you are)

Clay Barker, PhD







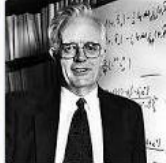

JMP Principal Research Statistician Developer

In 1996, Bradley Efron of Stanford was asked to name the most important problems in statistics.

most influential statisticians

All News Images Videos Shopping More Settings Tools

Statisticians

 <p>George E. P. Box 1919–2013</p>	 <p>John Tukey 1915–2000</p>	 <p>Ronald Fisher 1890–1962</p>	 <p>Karl Pearson 1857–1936</p>	 <p>Gertrude Mary Cox 1900–1978</p>	 <p>Bradley Efron</p>	 <p>David Cox</p>	 <p>Walter A. Shewhart 1891–1967</p>
---	---	--	---	---	--	--	---

# Variable Selection is a Big Deal

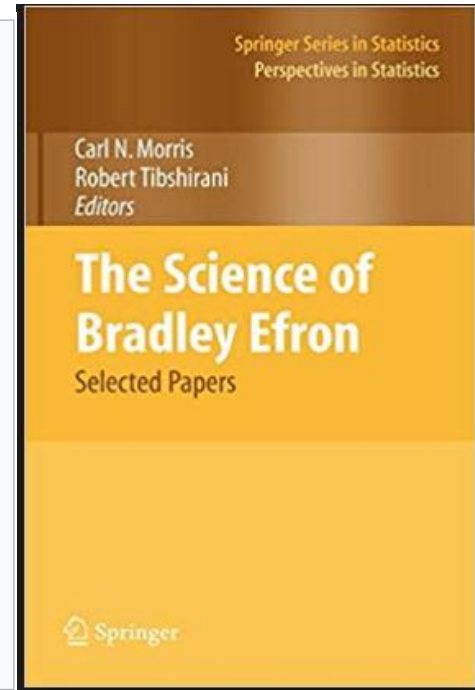
Efron's response was a single topic:  
**Variable selection in regression**

Paraphrased from an excellent  
survey paper:

Hesterberg et al. (2008),

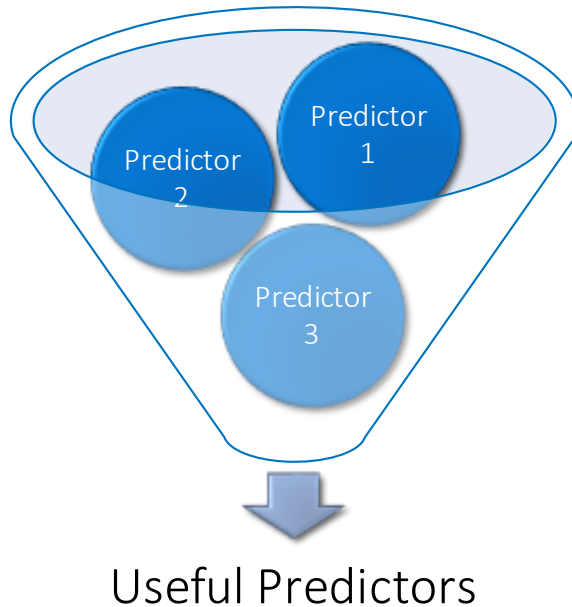
“Least angle and L1 penalized  
regression: A Review”, *Statistics  
Surveys*, 2, 61-93.

	<b>Bradley Efron</b>
<b>Born</b>	May 24, 1938 (age 81)
<b>Nationality</b>	American
<b>Alma mater</b>	California Institute of Technology, Stanford University
<b>Known for</b>	Bootstrap method
<b>Awards</b>	National Medal of Science (2005) BBVA Foundation Frontiers of Knowledge Award (2016) International Prize in Statistics (2019)
	<b>Scientific career</b>
<b>Fields</b>	Statistics
<b>Institutions</b>	Stanford University
<b>Thesis</b>	<i>Problems in Probability of a Geometric Nature</i> (1964)
<b>Doctoral advisor</b>	Rupert Miller Herbert Solomon <sup>[citation needed]</sup>
<b>Doctoral students</b>	Norman Breslow Robert Tibshirani



# 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 mean we don't have to think about what we're doing!

# What is Variable Selection?

Variable selection is crucial for a variety of reasons.

The resulting model...

1. ...is easier to interpret. Often it is much easier to interpret
2. ...will generalize well to new observations.
3. ...is stable to small changes in the observed data.
4. ...is easier to use/deploy.

It goes by several names: predictor/feature/subset selection and others

# Variable Selection

## In our daily lives

We run into a similar situation in our own decision making process.

## Why Too Much Data Disables Your Decision Making

Best known for killing cats, curiosity can also slay your judgment.

Posted Dec 04, 2012

From Psychology Today

## Information overload is killing our ability to make decisions

From Business Insider



Rikke Duus and Mike Cooray, The Conversation Jul 15, 2015, 10:00 PM



# An Example

## Reactor 32 Runs.jmp

From Box, Hunter, and Hunter (1978) and available in JMP's sample data.  
Want to understand a chemical reaction based on several factors.

	Pattern	Feed Rate	Catalyst	Stir Rate	Temperature	Concentration	Percent Reacted
1	-----+	10	1	100	140	6	56
2	----+-	10	1	100	180	3	69
3	---+--	10	1	120	140	3	53
4	-----+	10	1	120	180	6	49
5	---+--	10	2	100	140	3	63
6	---+--	10	2	100	180	6	78
7	---+--	10	2	120	140	6	67
8	---+--	10	2	120	180	3	95
9	++++--	15	1	100	140	3	53
10	++++--	15	1	100	180	6	45

# Reactor Example

5 main effects + interactions = 15 effects to consider

Why do we need to do variable selection?

1. It's unlikely that all 15 effects impact the response
2. If our goal is interpretation, we'd like a simple explanation.
3. If our goal is prediction, we want accurate predictions.
4. You should always do variable selection\*

\* My opinion 😊



# Reactor Example

DEMO

# Reactor Example

## What did we learn?

Even for a small example, trying to manually build a model is problematic.

...and once we're done, our model may not be great.

...and for larger problems, a manual process is probably not feasible.

We need a more structured approach to model building.

# Outline

1. A Brief Introduction to the Generalized Regression platform
2. A Brief Introduction to Variable Selection Techniques
  1. Step based methods
  2. Penalized regression
3. Some interesting use cases
  1. Functional Data
  2. Censoring
  3. Non-continuous response variable



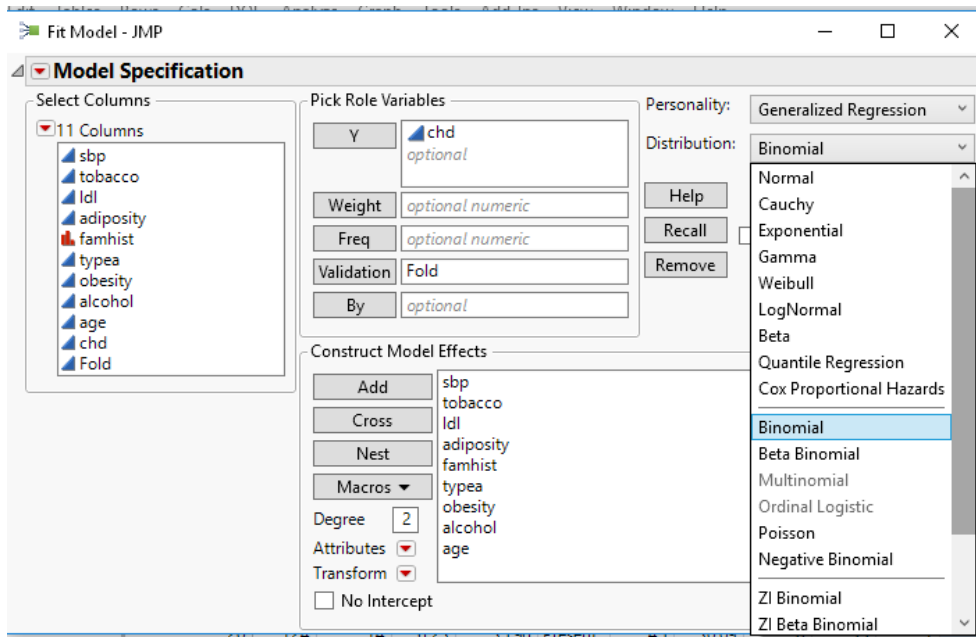
# The Generalized Regression Platform

A Brief Introduction

# The Generalized Regression Platform

## What is it?

Fit Model personality introduced in JMP Pro 11. We like to call it Genreg.



# Genreg

## Response Distributions

Genreg can handle a wide variety of response types since we can't always assume that our response is normally distributed.

Count data, skewed responses, labels, outliers...

We won't have time to cover the details of GLMs today, but...

Session ID: 2019-US-45MP-183

### Not Quite Normal: Choosing the Best Distribution for Modeling Your Response

Clay Barker, JMP Principal Research Statistician Developer, SAS

TOPIC: PREDICTIVE MODELING

LEVEL: ✨ ✨

Thursday 1:45-2:30  
Indigo

# Genreg

## Estimation and Selection

Genreg has a variety of estimation/selection methods to choose from

- Maximum Likelihood: full fit with no variable selection
- Step based methods – Forward, Backward, Best subset,...
- Penalized methods – Lasso, Elastic Net, Dantzig Selector,...

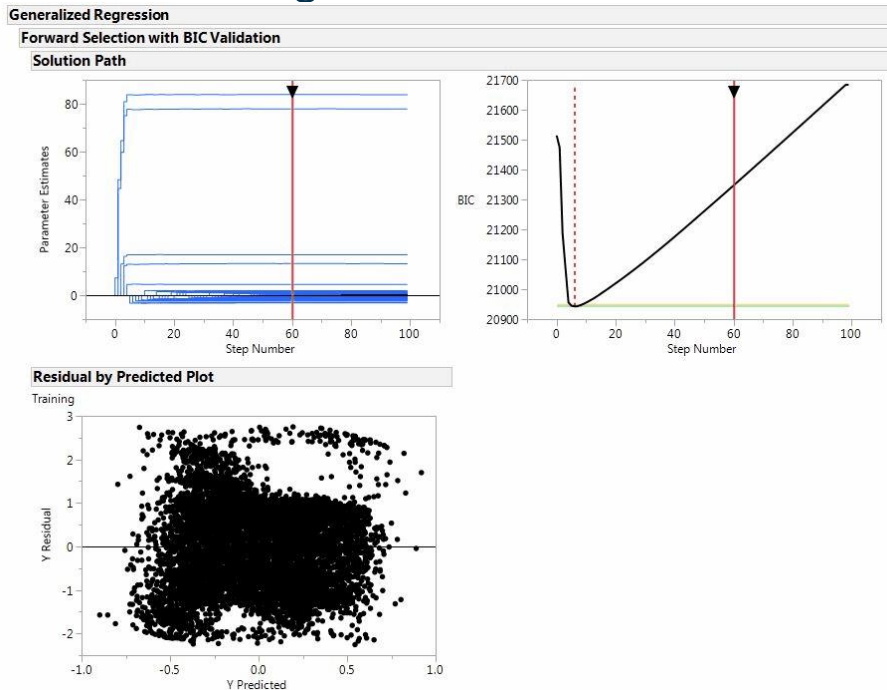
And a variety of validation methods to tune these methods

- Information based (AIC, BIC, ERIC)
- Cross-validation (k-fold, holdback, ...)

# Genreg

## Interactivity

The interactivity makes building models easier. More on this soon.

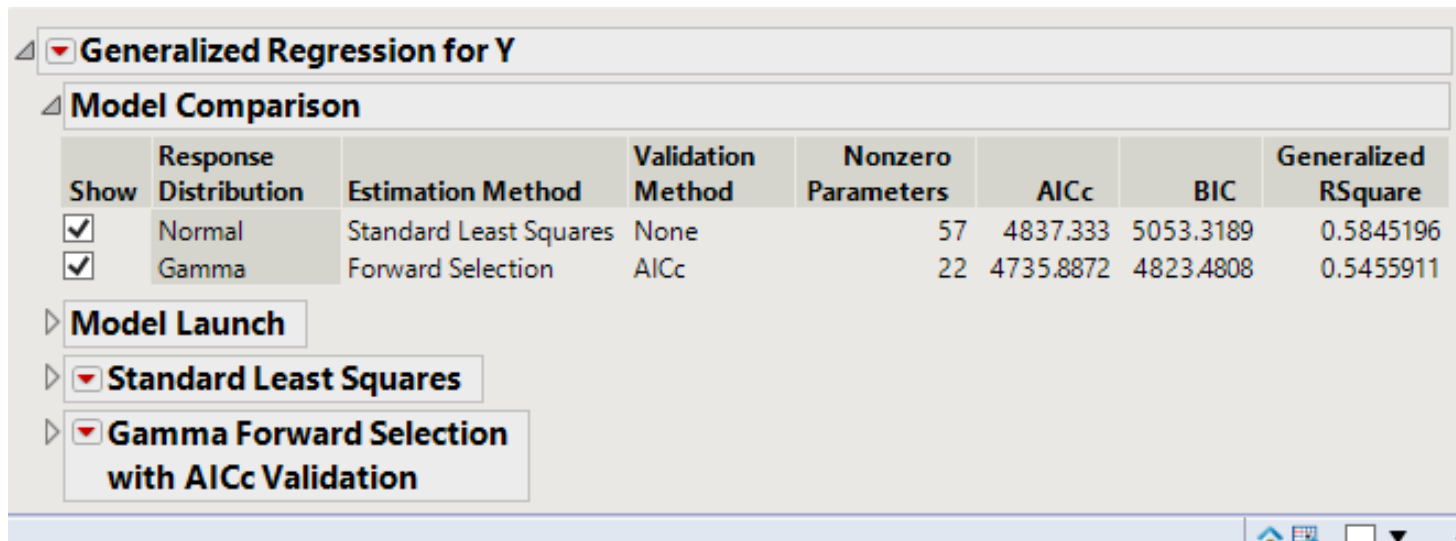




# Genreg

## What's New in 15?

Most substantial new feature is probably the ability to more easily compare models within a single platform launch.



The screenshot displays the SAS Genreg software interface. At the top, there is a dropdown menu labeled "Generalized Regression for Y". Below it, a section titled "Model Comparison" contains a table with the following data:

Show	Response Distribution	Estimation Method	Validation Method	Nonzero Parameters	AICc	BIC	Generalized RSquare
<input checked="" type="checkbox"/>	Normal	Standard Least Squares	None	57	4837.333	5053.3189	0.5845196
<input checked="" type="checkbox"/>	Gamma	Forward Selection	AICc	22	4735.8872	4823.4808	0.5455911

Below the table, there are three expandable sections: "Model Launch", "Standard Least Squares", and "Gamma Forward Selection with AICc Validation". The "Gamma Forward Selection with AICc Validation" section is currently expanded. At the bottom right of the interface, there are navigation icons for home, print, and a dropdown arrow.

# Genreg

## Our Mission

Genreg's goal: One framework to interactively build regression models.

- ...regardless of what type of response you have – binary, time-to-event,...
- ...whether you're analyzing the results of a designed experiment or an observational study.

Genreg can be your go-to place to build regression models in JMP Pro.

Before getting to variable selection methods, let's look at the interactive solution path.



# The Interactive Solution Path

# The Interactive Solution Path

Familiarity with the selection methods in Genreg is important.  
But understanding the solution path is also key.

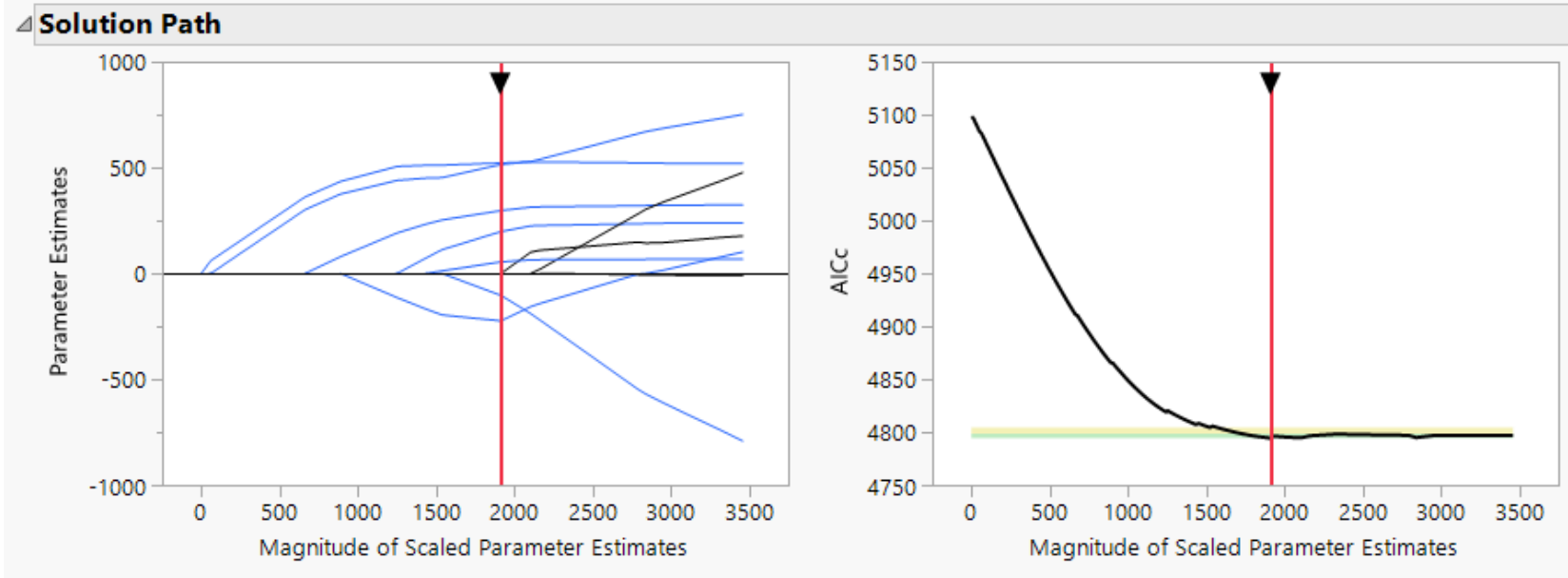
Most (all?) variable selection techniques can be described as two steps:

1. Fit a sequence of models
2. Pick the best model in the sequence based on some criteria

Those two steps can be conveniently summarized in the solution path plot.

# The Interactive Solution Path

## What is it?

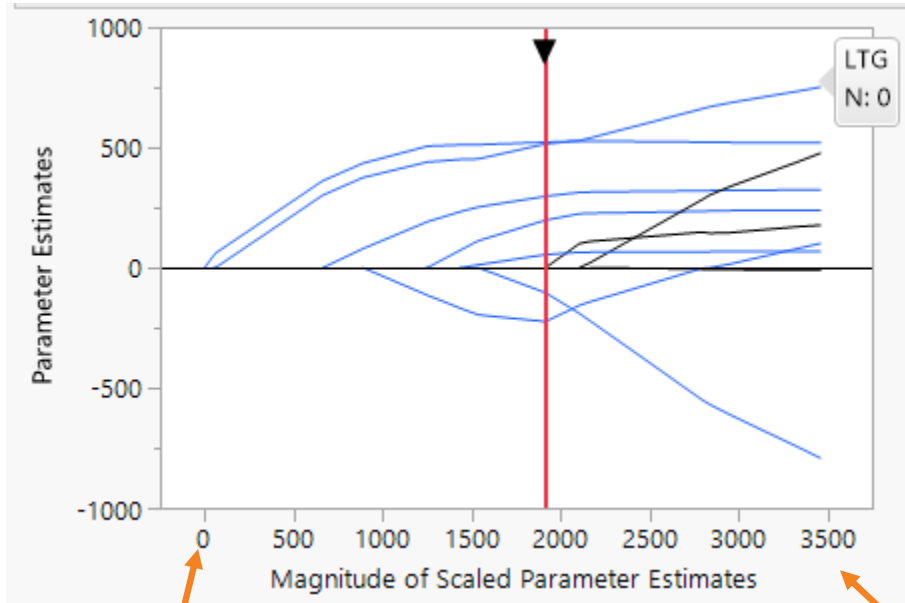


Summary of how the model changes along the sequence

Summary of how well the model fits along the way

# The Interactive Solution Path

## A closer look



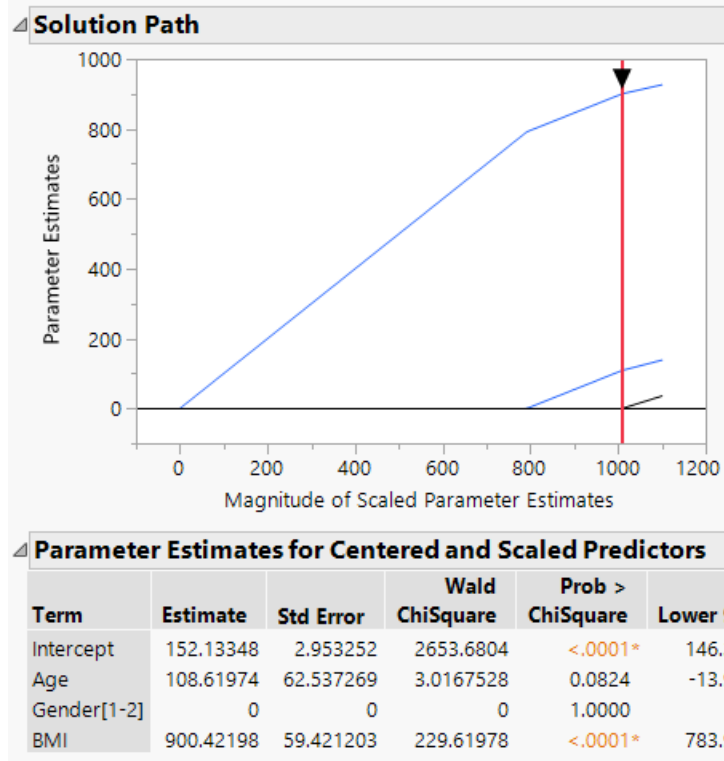
Least complex  
(just an intercept)

Most complex  
(all of the predictors)

- Each trace represents a regression parameter's estimate as you move through the sequence.
- When a parameter estimate is zero, the selection process has removed it from the model.

# The Interactive Solution Path

## What is the horizontal axis?

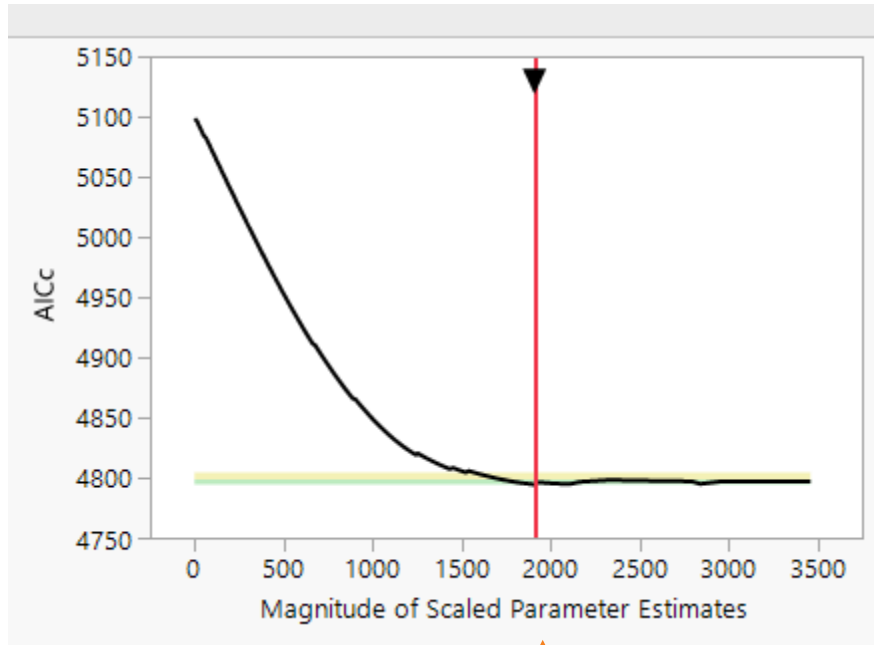


- The horizontal axis is a measure of the complexity of the model.
- Here, it is the  $\ell_1$  norm of the parameter estimates.

$$= \sum_{j=1}^p |\hat{\beta}_j|$$

- In this case, the red line is at  $108.6 + 0 + 900.4 \approx 1009$
- For other methods, we may use the step number in the algorithm.

# The Interactive Solution Path



- This piece tells us how well each model in the sequence fits.
- AICc is an information criteria, smaller is better.
- Same idea as if we had used cross-validation to pick the best.



Models get better as they get more complex up until here, where they start to get worse.



# The Interactive Solution Path

## How well does a model fit?

There are many ways to evaluate how well a model fits...and unfortunately we don't have time to cover this topic today.

\*But don't stress, I have other slides on the community about validation.

But here's what Genreg offers in a nutshell:

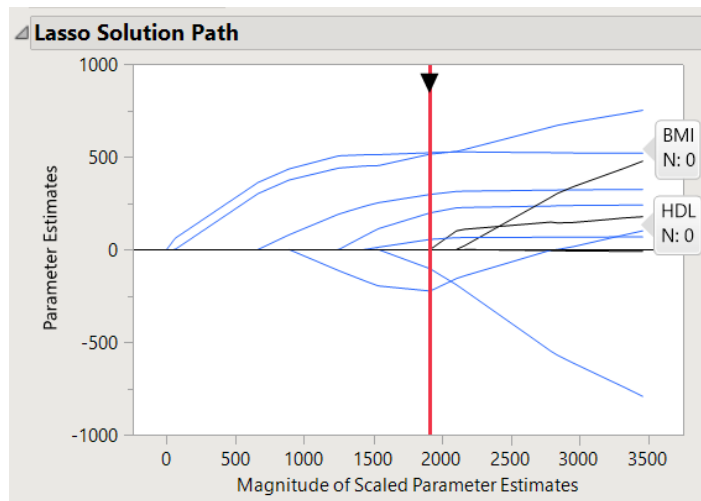
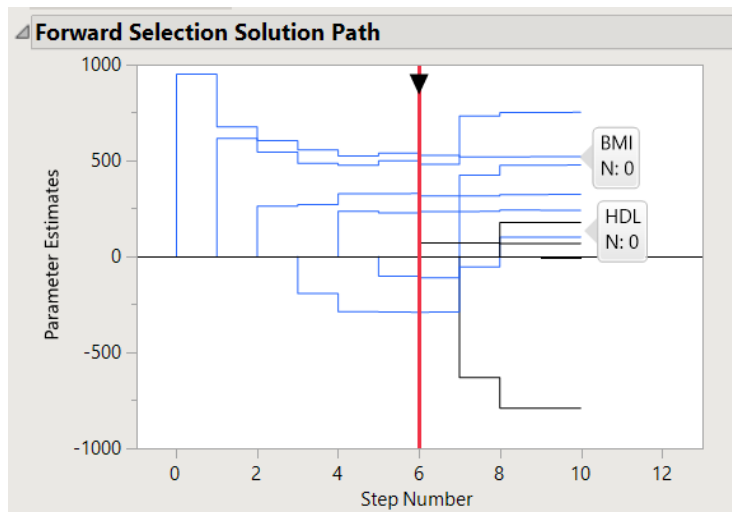
1. Information Criteria – a measure that balances how well the model fits the data with how many parameters it contains. Ex: AICc and BIC
2. Cross-validation – use one set of data to fit the model and a second set of data to evaluate how well it will fit on new data. Ex: holdback and k-fold.

Which one to use?

Information criteria for small data, CV for more data is a decent rule.

# Interactive Solution Path

## Stepwise vs Penalized



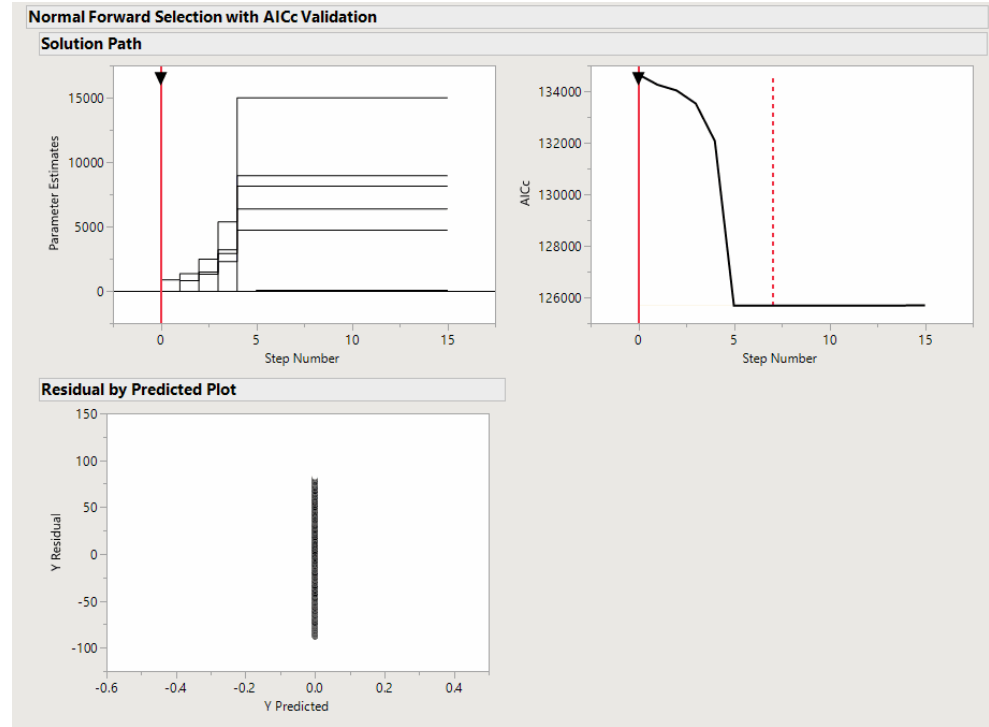
The shapes may be different, but the information conveyed is the same regardless of method: The sequence of variables selected.

# Interactive Solution Path

## Brace yourself for a bunch of gifs

If you're not terribly comfortable with the variable selection methods in Genreg, it's not the end of the world.

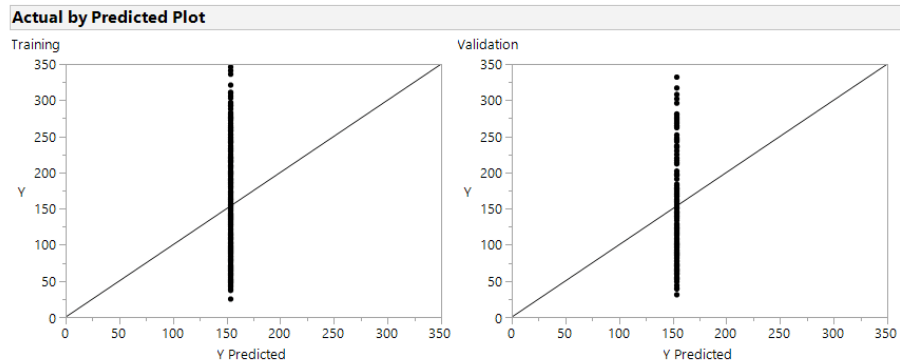
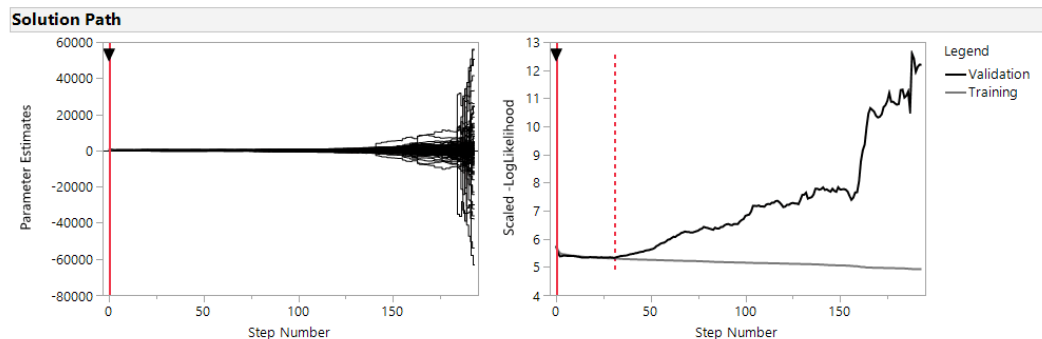
Understanding the path and taking advantage of its interactivity is usually sufficient



# Interactive Solution Path And Diagnostics

Taking advantage of interactivity and built-in diagnostics help us understand our fits.

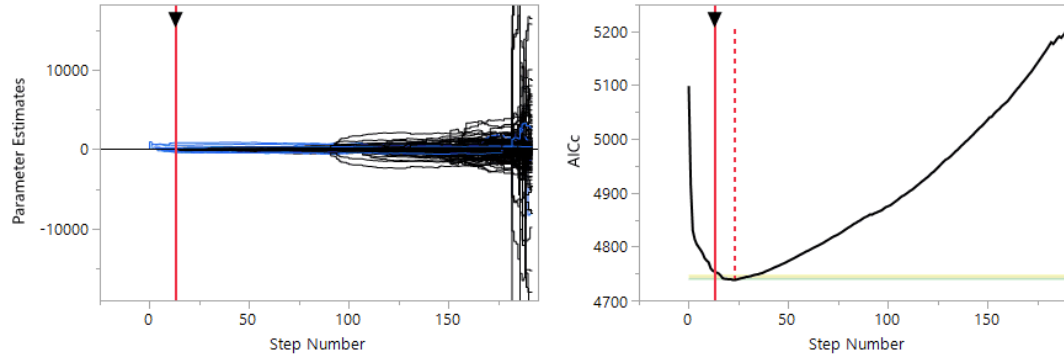
Our model slows down improving on training and starts to get worse on the hold-out set.



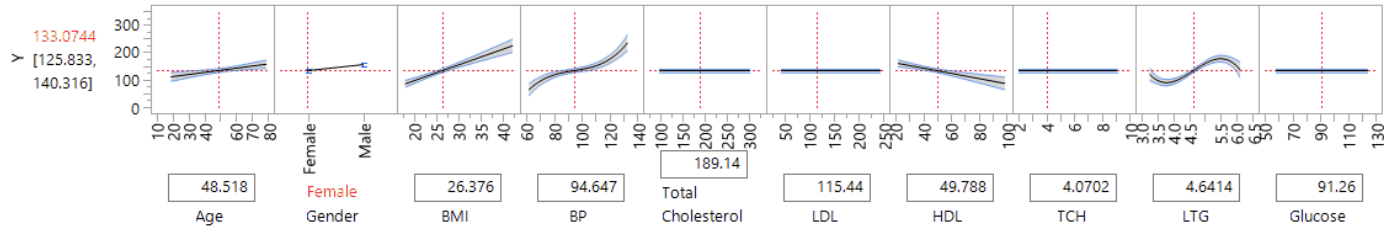
# Interactive Solution Path

## Overfitting and the Profiler

**Solution Path**



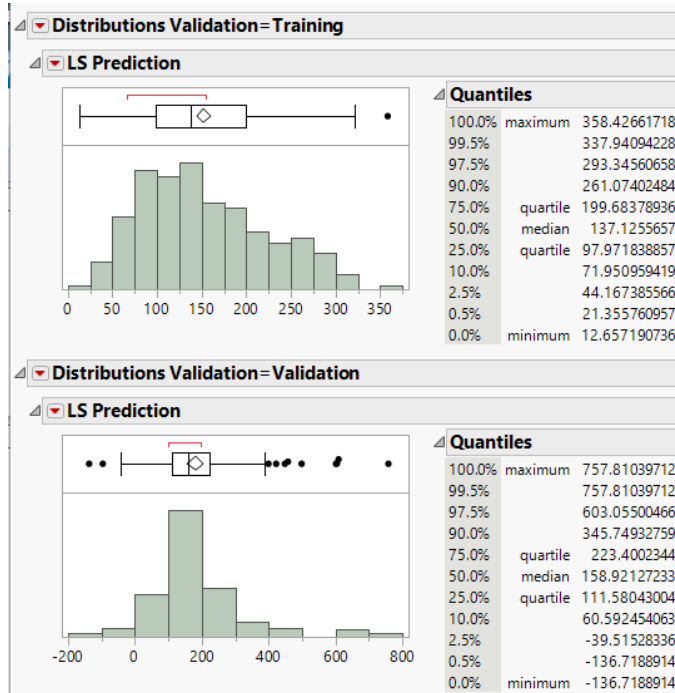
**Prediction Profiler**



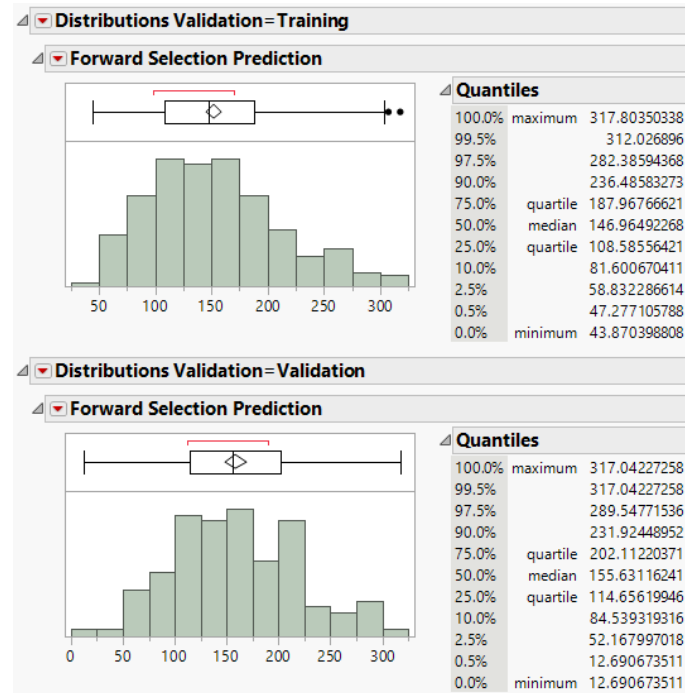
# Interactive Solution Path

## What does that variance mean for predictions?

An Overfit Model



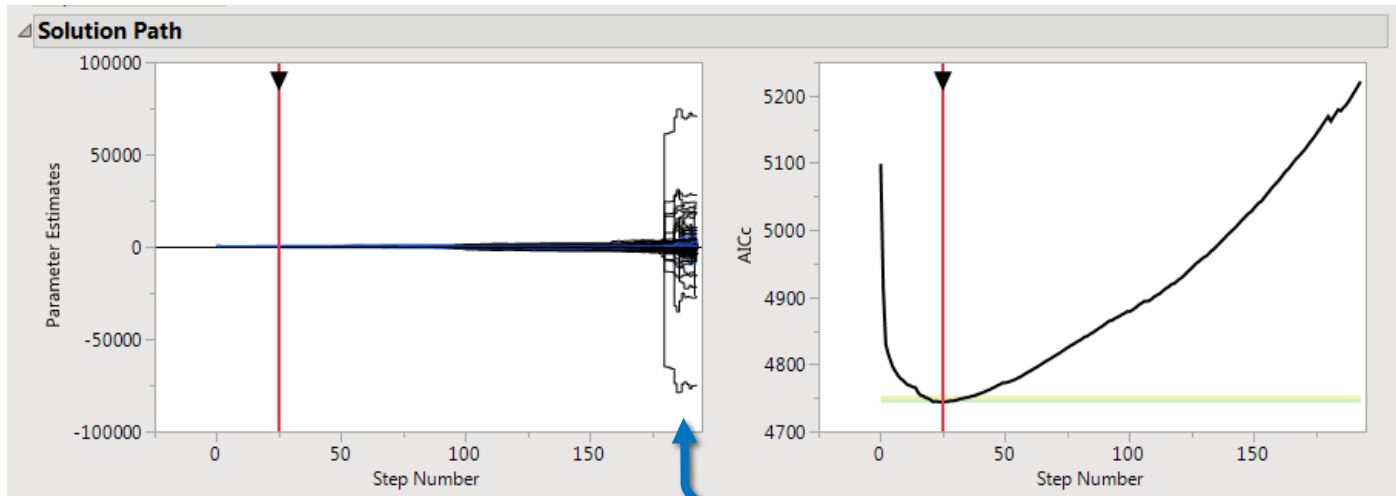
A More Appropriate Model



# Interactive Solution Path

## Collinearity

As we add more collinearity to our model, things get unstable and our estimates blow up in magnitude towards the end of the path.

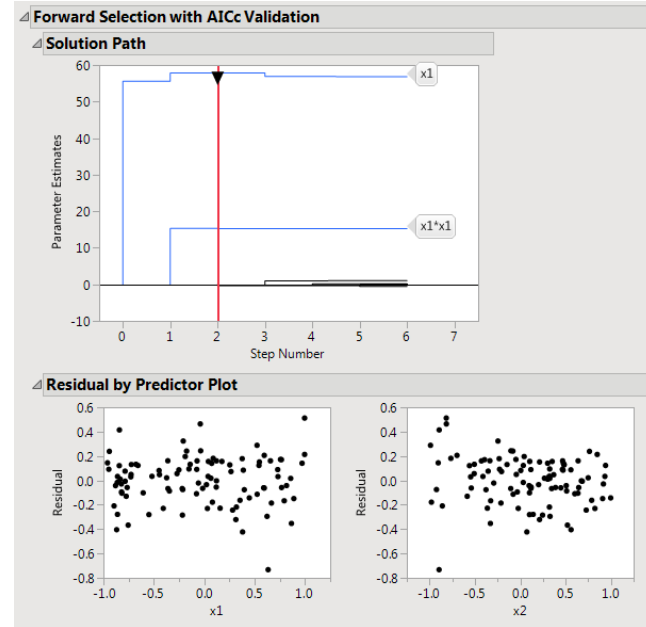
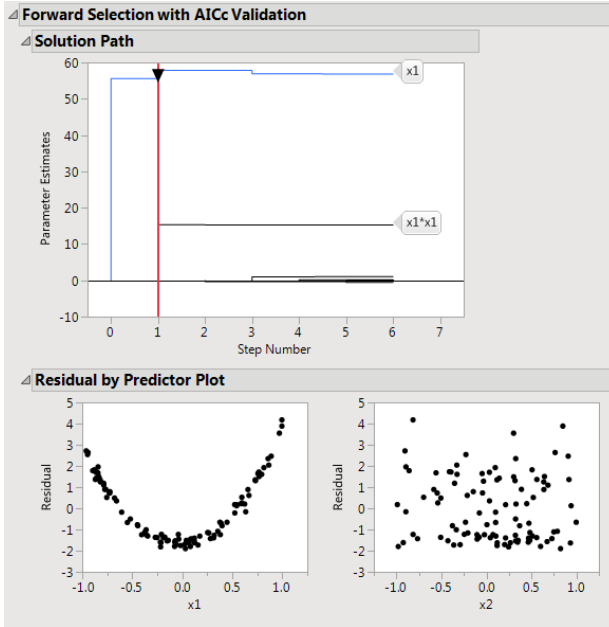


Instability

# Interactive Solution Path

## Missing Effects

- Keep an eye on residual plots – you may realize you're missing an effect.







# Step-based Methods in Genreg

# Stepwise Methods in Genreg

Genreg offers a handful of *stepwise* variable selection methods.

Why do we call them stepwise?

- They are largely algorithmic methods.
- Given the current model, how do we improve our model in the next step by adding or removing a variable?

Stepwise methods in Genreg

1. Best Subset
2. Forward Selection and Two-Stage Forward Selection
3. Backward Elimination
4. Pruned Forward Selection

# 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 best based on a validation method.

Here's a simple case with three candidate predictors (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)

So we fit all 8 of these models and declare a best model based on some criterion.

That's great, right?

# Stepwise Methods in Genreg

## Best Subset

Sounds great!

We feel good because we know that we found the best of all models.

Now let's consider a bigger model.

Let's say that we have 10 continuous predictors.

We also want to consider two factor interactions and quadratic terms.

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

# Stepwise Methods in Genreg

## Best subset

How many models do we need to fit?

Number of terms	Number of Models	Running Total
1	65	66
2	2080	2146
3	43680	45826
4	677040	722866
5	8259888	8982754
6	82598880	91581634
7	696190560	787772194

We're only up to models of size 7 and we already have to fit almost 800M models...

# Stepwise Methods in Genreg

## Best Subset

- The number of models we need to fit explodes exponentially in  $p$ , the number of predictors.
- For our 65 predictor example, we have to fit

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

or almost 37 Quintillion models

just to say we have the best model using these 65 predictors!

- There are shortcuts, but best to stick to small problems.

## Best Subset Demo

# Stepwise Methods in Genreg

## Forward Selection

- Best Subset is not feasible for even moderately sized problems.
- 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 (Score). Add the variable with the best 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 predictors.

	<b>Step 1</b>	<b>Step 2</b>	<b>Step 3</b>	<b>Step 4</b>
<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*	

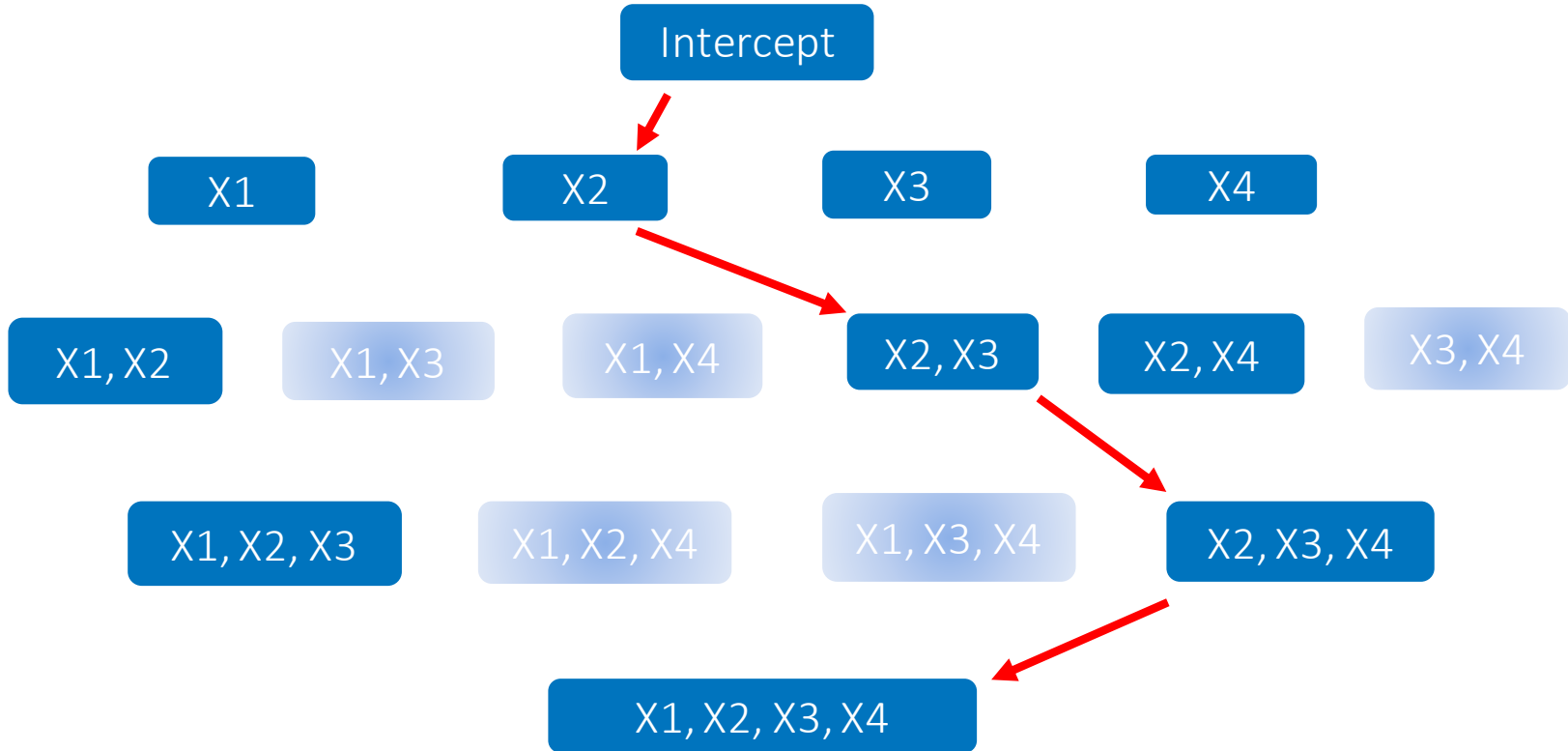
Now we have 5 models to fit and consider (out of 16 possible models)

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



# Stepwise Methods in Genreg

## Forward Selection

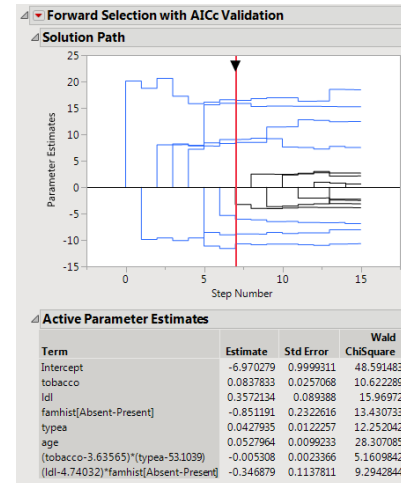
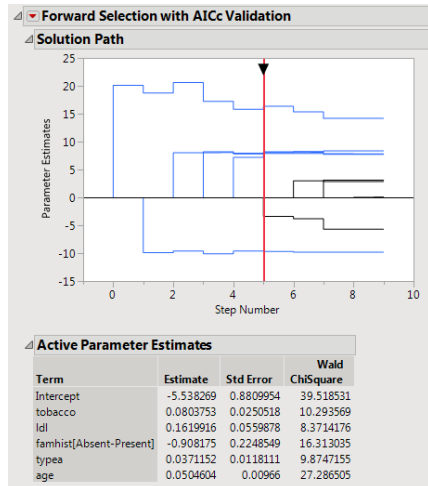


# Stepwise Methods in Genreg

## 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

## Backward Elimination

Backward Elimination puts structure around a manual process:

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

BE algorithm

1. Start with everything in the model
2. Drop the worst effect based on Wald 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.

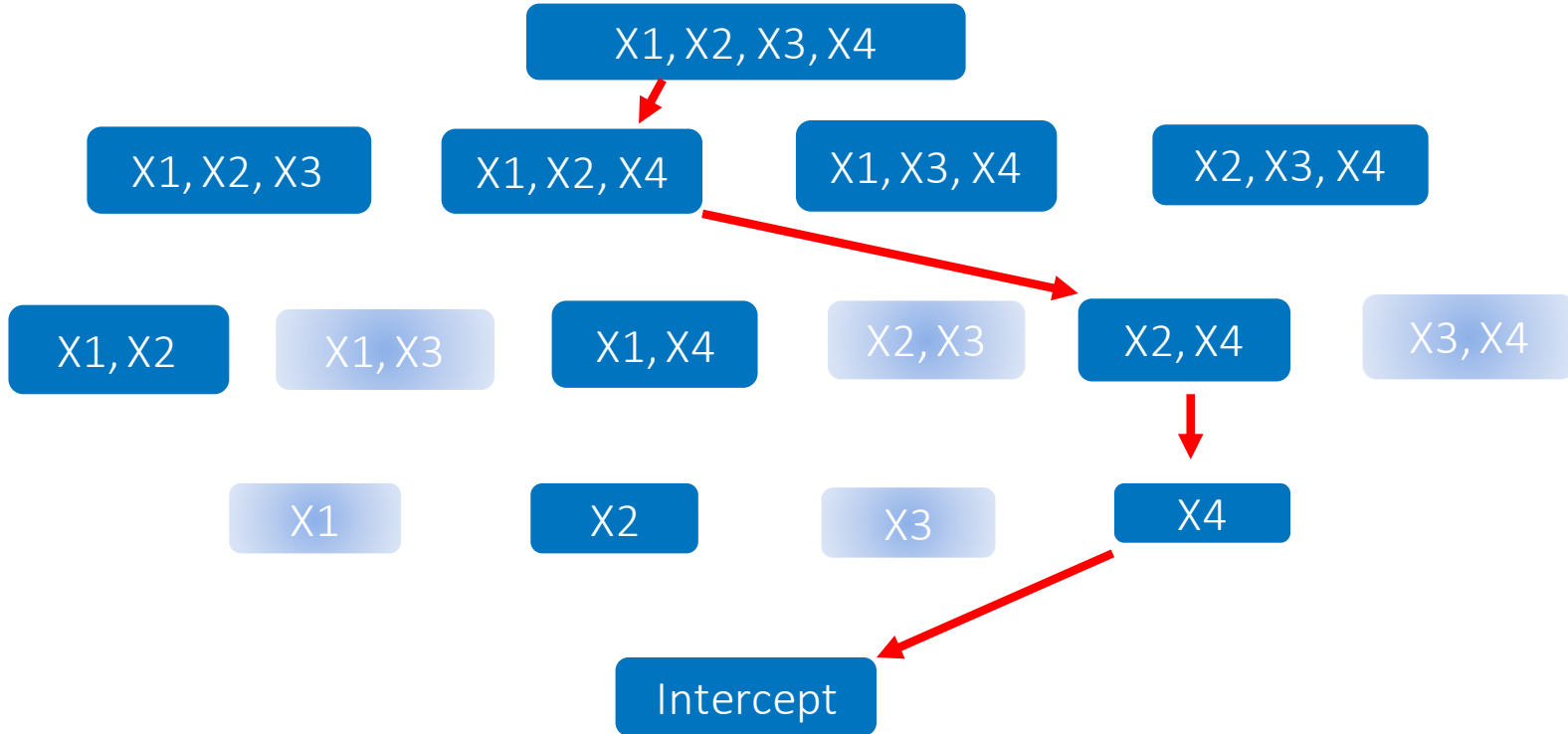
	<b>Step 1</b>	<b>Step 2</b>	<b>Step 3</b>	<b>Step 4</b>
$p$ for X1	.001	.4*		
$p$ for X2	.2	.15	.06*	
$p$ for X3	.6*			
$p$ for X4	.05	.03	.01	.003

Again we have a sequence of 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

## Backward Elimination



# Stepwise Methods in Genreg

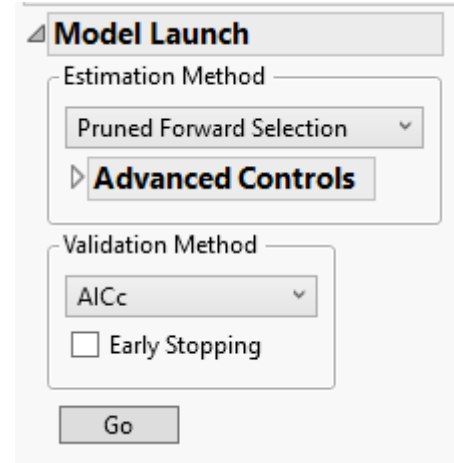
## Forward and Backward Steps

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 with 'Pruned Forward Selection' selected and a right-pointing arrow icon. Below this is a button labeled 'Advanced Controls'. The 'Validation Method' section contains a dropdown menu with 'AICc' selected and a right-pointing arrow icon, followed by a checkbox labeled 'Early Stopping' which is currently unchecked. At the bottom of the dialog 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

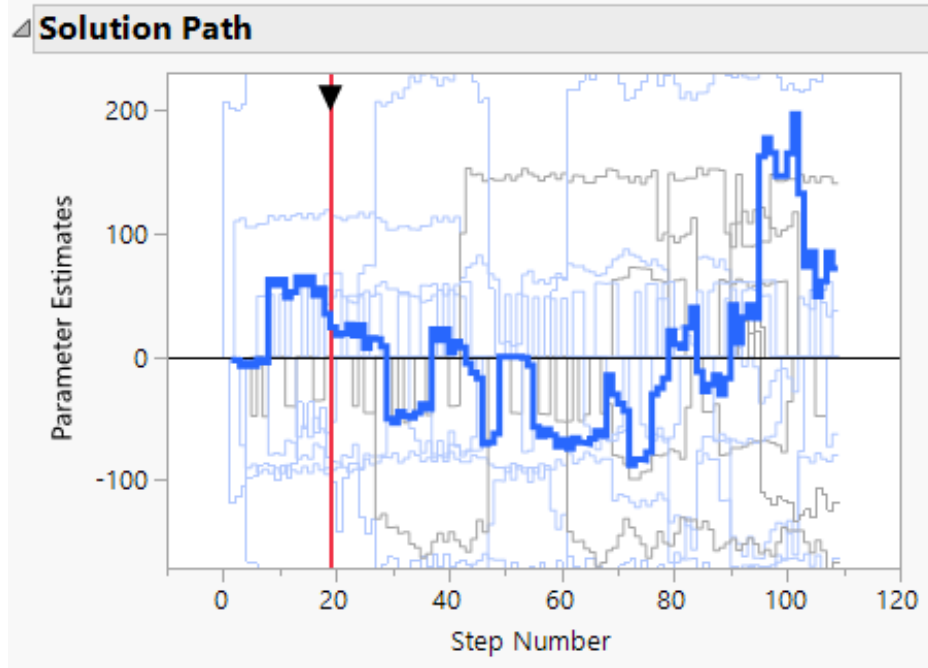
1. Find the variable that most wants to enter  $X_E$  (Score test)
2. Find the variable that most wants to leave  $X_L$  (Wald test)
  - A. Try adding  $X_E$
  - B. Try removing  $X_L$
  - C. Try swapping  $X_L$  for  $X_E$
3. Go with A, B, or C based on which fits best.
4. Go back to 1.

Starts like FS but then we prune off variables as we go.

# Stepwise Methods in Genreg

## Pruned Forward Selection

A variable can enter and leave the model many times (and change signs).

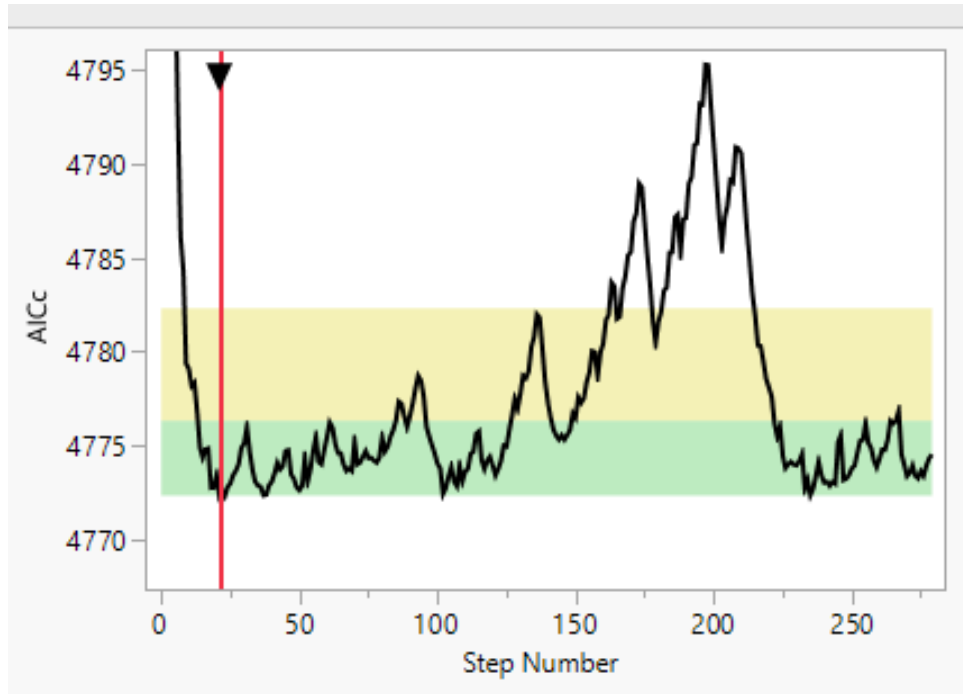




# Stepwise Methods in Genreg

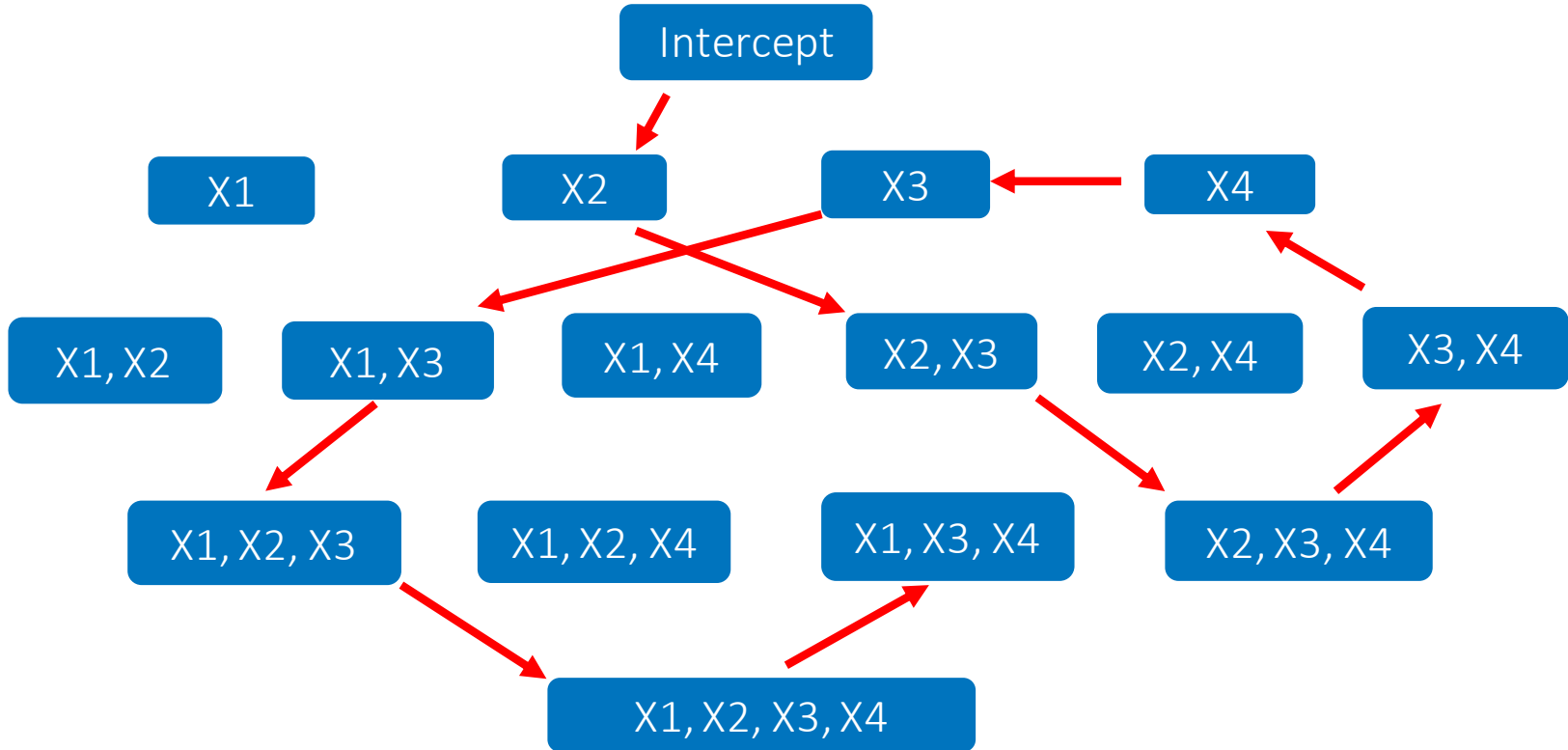
## Pruned Forward Selection

And the quality of the fit can vary wildly, usually a sign of collinearity.



# Stepwise Methods in Genreg

PFS bounces around a lot



# Stepwise Methods in Genreg

## Effect Heredity

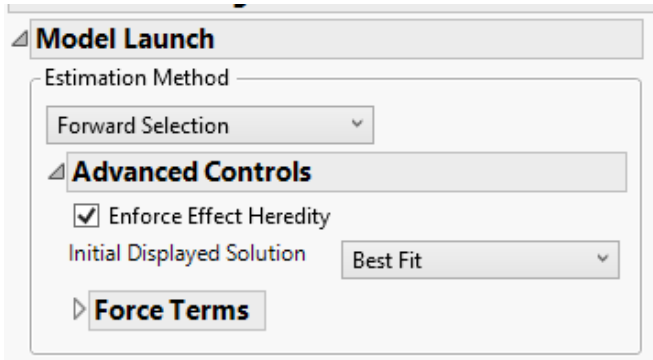
### Effect Heredity

If a higher order effect (interaction, quadratic, ...) is in the model, the lower order effects that compose it must also be in the model.

EX: We can't consider adding  $X^3$  to our model unless  $X$  and  $X^2$  are in.

EX: If we want to consider  $A * B * C$ ,

$A$ ,  $B$ ,  $C$ ,  $A * B$ ,  $A * C$ , and  $B * C$  must all be in the model



Heredity is enforced by default if we know that the data table is the result of a Designed Experiment.

Works naturally with stepwise methods.



# Penalized Regression Methods

Shrinkage and Selection

# Penalized Regression

Stepwise methods are great.

1. Easy to implement
2. Intuitive and easy to explain

But there are other ways to do variable selection.

Lots of interest recently in penalized regression methods because they do variable selection and shrinkage – both of which help to avoid overfitting.

Estimation tends to be more stable than searching (as in stepwise).

# Penalized Regression

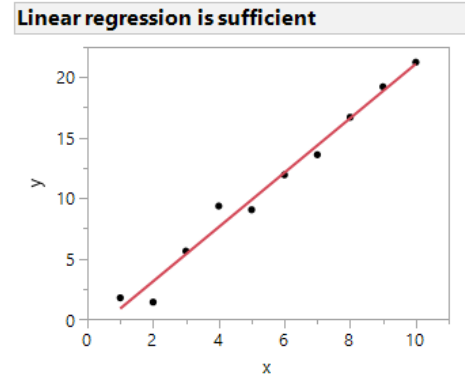
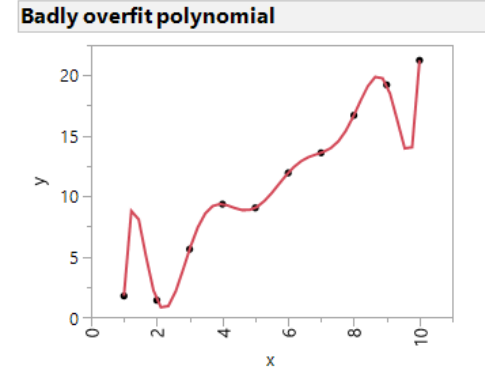
## Overfitting

What exactly is overfitting?

Overfitting occurs when our model is more complex than needed and it starts to model random noise in the data instead of the underlying relationships.

Classic overfitting

- Our model fits great on the observed data 😊
- Our model fails miserably when predicting new observations 😞
- Our inferences are misleading 😞
- If we slightly alter the data, our model may change dramatically 😞



# Penalized Regression

## Overfitting Example

We want to use information like age, gender, cholesterol, ...  
to predict an individual's diabetes progression over the course of a year.

$$E(Y_i) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Gender}_i + \beta_3 \text{BMI}_i + \dots$$

	Y	Age	Gender	BMI	BP	Total Cholesterol	LDL	HDL	TCH	LTG	Glucose
1	151	59	2	32.1	101	157	93.2	38	4	4.8598	87
2	75	48	1	21.6	87	183	103.2	70	3	3.8918	69
3	141	72	2	30.5	93	156	93.6	41	4	4.6728	85
4	206	24	1	25.3	84	198	131.4	40	5	4.8903	89
5	135	50	1	23.0	101	192	125.4	52	4	4.2905	80
6	97	23	1	22.6	89	139	64.8	61	2	4.1897	68
7	138	36	2	22.0	90	160	99.6	50	3	3.9512	82
8	63	66	2	26.2	114	255	185.0	56	4.55	4.2485	92
9	110	60	2	32.1	83	179	119.4	42	4	4.4773	94
10	210	70	1	30.0	85	180	102.4	42	4	5.2815	88

# Penalized Regression

## Overfitting

Break our data up: one piece for estimating our model and another to evaluate how our model fits on new data.

- 10 main effects + interactions + quadratics = 64 total terms
- If we use everything in our model, we overfit badly.

Measures of Fit for Y					
Validation	Predictor	RSquare	RASE	AAE	Freq
Training	OLS Pred	0.6645	44.537	34.933	265
Test	OLS Pred	0.1378	71.536	53.371	66



# Penalized Regression

## Overfitting

What if we did Forward Selection?

Our model would only have 7 terms and we'd predict new data better.

Measures of Fit for Y					
Validation	Predictor	RSquare	RASE	AAE	Freq
Training	OLS Pred	0.6645	44.537	34.933	265
Training	FS Pred	0.4908	54.866	45.405	265
Test	OLS Pred	0.1378	71.536	53.371	66
Test	FS Pred	0.4527	56.996	47.734	66

But can we do better still?

# Penalized Regression

## Prediction Error

Before we get to penalized methods, let's talk about prediction error.

Suppose we observe data of the form

$$Y_i = f(X_i) + \epsilon_i \quad i = 1, \dots, n$$
$$\epsilon \sim N(0, \sigma^2) \quad X_i \text{ is } p \times 1 \text{ vector of predictors}$$

$\hat{f}(X_i)$  is our fitted model.

We need a measure of how well we will predict a new observation:

$$\text{Prediction Error}(\hat{f}(X_{n+1})) = E\{[y_{n+1} - \hat{f}(X_{n+1})]^2\}$$

# Penalized Regression

## The Bias/Variance Tradeoff

$$\begin{aligned} \text{Prediction Error}(\hat{f}(X_{n+1})) &= \sigma^2 + \text{MSE}(\hat{f}(X_{n+1})) \\ &= \sigma^2 + \text{E}[f(X_{n+1}) - \hat{f}(X_{n+1})]^2 + \text{var}[\hat{f}(X_{n+1})] \end{aligned}$$

Fixed                      Bias Squared                      Variance

This is the bias/variance tradeoff in estimation.

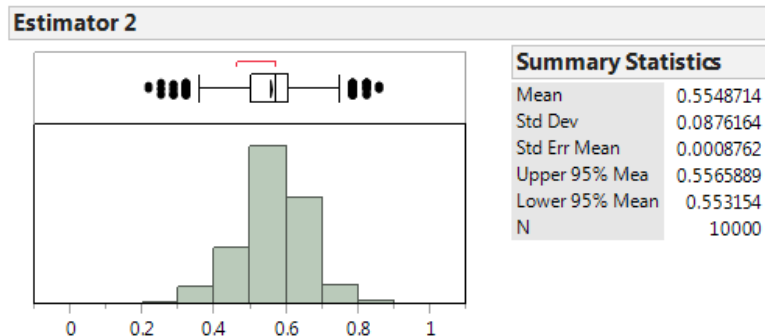
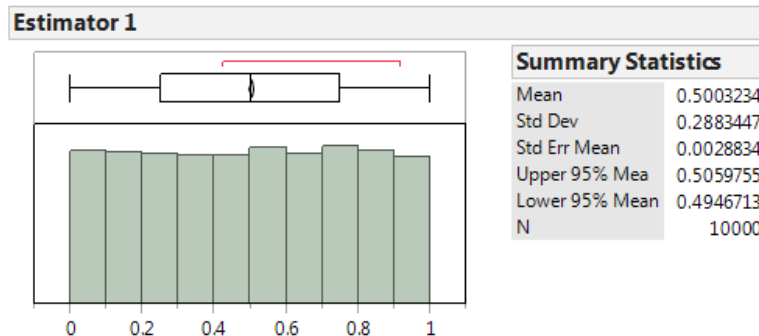
Maybe we can accept some bias to reduce variance?

This is the motivation behind penalized regression.

# Penalized Regression

## An exaggerated example of bias/variance tradeoff

- The estimator on the top is unbiased but highly variable.
- The estimator on the bottom is biased, but much less variable.













# Bias Variance Tradeoff

## A connection to planning my flights

### Option 1

### Option 2

3:22pm - 11:55pm 5h 33m (1 stop)    rc  
 American Airlines TUS - 44m in DFW - RDU  
Very Good Flight (8.3/10)  
[Details & baggage fees](#) 

3:22pm - 12:06am <sup>+1</sup> 5h 44m (1 stop)    roui +  
 American Airlines TUS - 54m in DFW - RDU  
Very Good Flight (8.3/10)  
[Details & baggage fees](#) 

Roundtrip price for 1 traveler, including taxes  
**Free Cancellation** within :

Roundtrip price for 1 traveler, including taxes  
**Free Cancellation** within 24


3:22pm → 7:36pm 2h 14m  
Tucson to Dallas  
Tucson Intl. (TUS) to Dallas-Fort Worth Intl. (DFW)  
American Airlines 1097  
Economy / Coach (V)  
Boeing 737-800 | Food For Purchase

3:22pm → 7:36pm 2h 14m  
Tucson to Dallas  
Tucson Intl. (TUS) to Dallas-Fort Worth Intl. (DFW)  
American Airlines 1097  
Economy / Coach (V)  
Boeing 737-800 | Food For Purchase

 **44m stop** Dallas (DFW)

 **54m stop** Dallas (DFW)

8:20pm → 11:55pm 2h 35m  
Dallas to Raleigh  
Dallas-Fort Worth Intl. (DFW) to Raleigh - Durham Intl. (RDU)  
American Airlines 347  
Economy / Coach (V)  
Boeing 737-800 | Food For Purchase

8:30pm → 12:06am 2h 36m  Overnight - Arrives Sat, Oct 19  
Dallas to Raleigh  
Dallas-Fort Worth Intl. (DFW) to Raleigh - Durham Intl. (RDU)  
American Airlines 1155  
Economy / Coach (V)  
Boeing 737-800 | Food For Purchase

# Penalized Regression

## Ridge Regression

OLS is unbiased and we can work out the prediction error.

What if we minimize a penalized sum of squared errors instead?

$$\begin{aligned}\hat{\beta}_{ridge} &= \arg \min_{\beta} \sum_i (y_i - x_i \beta)^2 + \frac{\lambda}{2} \sum_j \beta_j^2 \\ &= (X^T X + \lambda I_p)^{-1} X^T y\end{aligned}$$

Tuning parameter  $\lambda$  controls the magnitude of parameters.

- $\lambda = 0$  is the usual OLS solution
- As  $\lambda$  increases, parameter estimates move toward zero. Shrinkage!

# Penalized Regression

## Ridge MSE

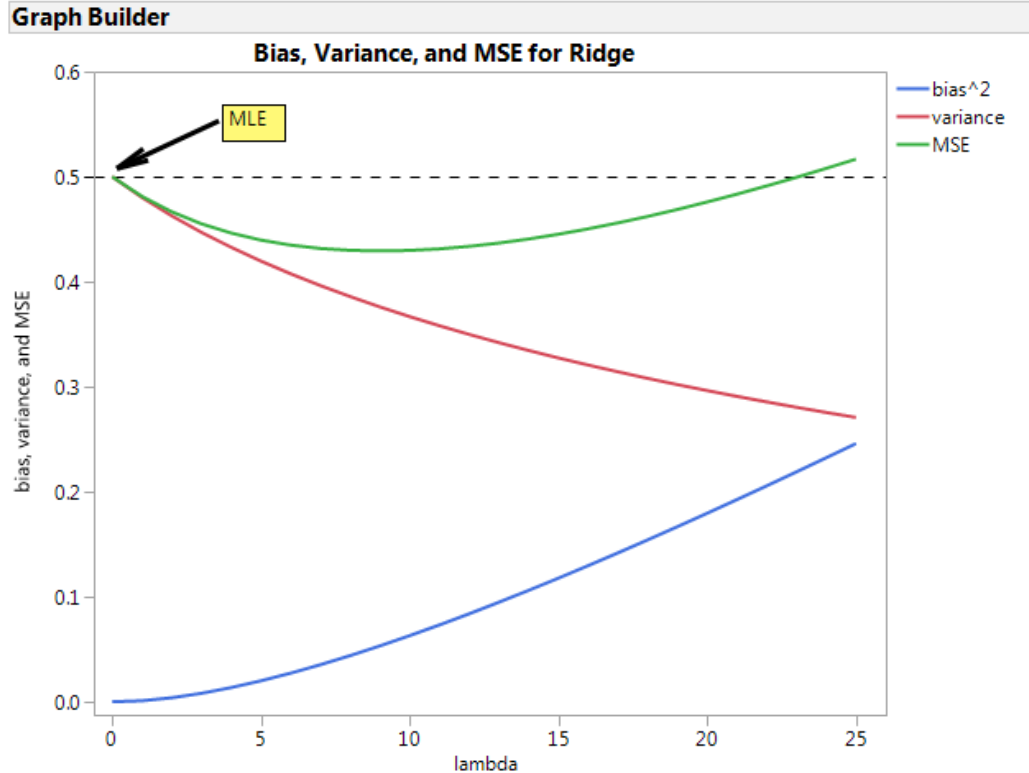
- $$\begin{aligned} E(\hat{\beta}_{ridge}) &= [I_p + \lambda(X^T X)^{-1}]^{-1} E(\hat{\beta}_{LS}) \\ &= [I_p + \lambda(X^T X)^{-1}]^{-1} \beta \end{aligned}$$
- $$\text{var}(x_i \hat{\beta}_{ridge}) = \sigma^2 x_i^T (X^T X + \lambda I_p)^{-1} [I_p + \lambda(X^T X)^{-1}]^{-1} x_i$$

  1. Ridge estimates are biased toward zero
  2. Ridge estimates are less variable than OLS.

The big question: Can Ridge MSE beat the OLS MSE?

# Penalized Regression

## Ridge vs OLS



- The answer:  
It depends on  $\lambda$
- $\lambda \in (0, 22]$  Ridge beats OLS, otherwise Ridge is worse
- This is a simulated example with  $N=100$  and  $p=50$ .



# Penalized Regression

## Ridge and the Diabetes data

Back to the Diabetes example. How does Ridge do?

Measures of Fit for Y								
Validation	Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq	
Training	OLS Pred	Fit Least Squares		0.6645	44.537	34.933	265	
Training	FS Pred	Fit Generalized Forward Selection		0.4908	54.866	45.405	265	
Training	Ridge Pred	Fit Generalized Ridge		0.5440	51.918	43.039	265	
Test	OLS Pred	Fit Least Squares		0.1378	71.536	53.371	66	
Test	FS Pred	Fit Generalized Forward Selection		0.4527	56.996	47.734	66	
Test	Ridge Pred	Fit Generalized Ridge		0.4893	55.054	45.817	66	

We do a good job predicting new observations, but remember that ridge regression *does not* do variable selection.

# Penalized Regression

## A Family of Models

Ridge opened the door to a variety of penalized techniques

$$\hat{\beta} = \arg \min_{\beta} \sum_i (y_i - x_i \beta)^2 + \lambda \sum_j \rho(\beta_j)$$

$\rho(x)$	Technique
$x^2$	Ridge (L2 norm)
$ x $	Lasso (L1 norm)
$I(x \neq 0)$	Best Subset (L0 norm)
$I(x \leq \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} I(x > \lambda)$	Smoothly clipped absolute deviation

We have no plans to implement SCAD in JMP, but the point is that there are many types of penalties out there.

# Penalized Regression

## The Lasso

Tibshirani (1996) introduced the Lasso:

$$\hat{\beta}_{lasso} = \arg \min_{\beta} \sum_i (y_i - x_i \beta)^2 + \lambda \sum_j |\beta_j|$$

- Biases coefficients by shrinking them toward zero, like ridge.
- Unlike ridge, it can shrink estimates all the way to zero. (selection)  
Least absolute shrinkage and selection operator
- The absolute value penalty is a pain compared to ridge.

# Penalized Regression

## Back to Diabetes

Unfortunately the lasso MSE is not easy to work out.

- As  $\lambda$  increases, more bias and less variance.
- If only a subset of predictors truly are active, lasso should beat ridge.

Lasso has a slight edge on Test set and it only includes 9 predictors.

Measures of Fit for Y					
Validation	Predictor	RSquare	RASE	AAE	Freq
Training	OLS Pred	0.6645	44.537	34.933	265
Training	Ridge Pred	0.5440	51.918	43.039	265
Training	Lasso Pred	0.5183	53.363	44.611	265
Test	OLS Pred	0.1378	71.536	53.371	66
Test	Ridge Pred	0.4893	55.054	45.817	66
Test	Lasso Pred	0.5085	54.010	45.781	66

# Penalized Regression

## Ridge vs Lasso

### Ridge

- Provides an estimate for all  $p$  terms (even when  $n < p$ )
- Naturally handles collinearity and even linear dependencies

### Lasso

- Estimation and variable selection at the same time
- Provides estimates for up to  $n$  parameters
- If  $x_1$  and  $x_2$  are highly correlated, we'll probably only select **one** of them.

Can we combine their strengths?

# Penalized Regression

## The Elastic Net

Zou and Hastie (2005): Ridge + Lasso = Elastic Net

$$\text{Penalty: } \rho(\beta) = \frac{1-\alpha}{2} \beta^2 + \alpha |\beta| \quad \alpha \in [0,1]$$

- $\alpha$  tuning parameter controls the mix of  $\ell_1$  and  $\ell_2$  penalties.
- Ridge and Lasso are special cases ( $\alpha = 0$  and  $\alpha = 1$  respectively)

When  $\alpha \in (0,1)$

1. We get selection and shrinkage
2. We can handle collinearity and dependencies.
3. We can estimate more than  $n$  coefficients.

Just stick with  $\alpha$  close to 1 (default is .99 in Genreg)

# Penalized Regression

## Elastic Net vs Lasso

Suppose we have 10 candidate predictors.

$x_2$  and  $x_4$  are highly correlated and at least one of them is truly active.

- Lasso will likely only choose  $x_2$  **or**  $x_4$
- Elastic Net will likely choose  $x_2$  **and**  $x_4$

Which solution is better? It depends on context.

Lasso often fits better, but elastic net's interpretation may be important.

Another distinction: Elastic net can select more than  $n$  parameters, which may or may not be a good thing.

# Penalized Regression

## Diabetes yet again!

Elastic Net does slightly better on the Test set than Lasso.

Elastic Net chooses 32 variables, Lasso only 9.

Why? Our variables are highly correlated (BMI, BP, Cholesterol,...)

Measures of Fit for Y					
Validation	Predictor	RSquare	RASE	AAE	Freq
Training	OLS Pred	0.6645	44.537	34.933	265
Training	Lasso Pred	0.5183	53.363	44.611	265
Training	Elastic Net Pred	0.5808	49.782	40.860	265
Test	OLS Pred	0.1378	71.536	53.371	66
Test	Lasso Pred	0.5085	54.010	45.781	66
Test	Elastic Net Pred	0.5296	52.838	42.988	66



# Penalized Regression

## The Dantzig Selector

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

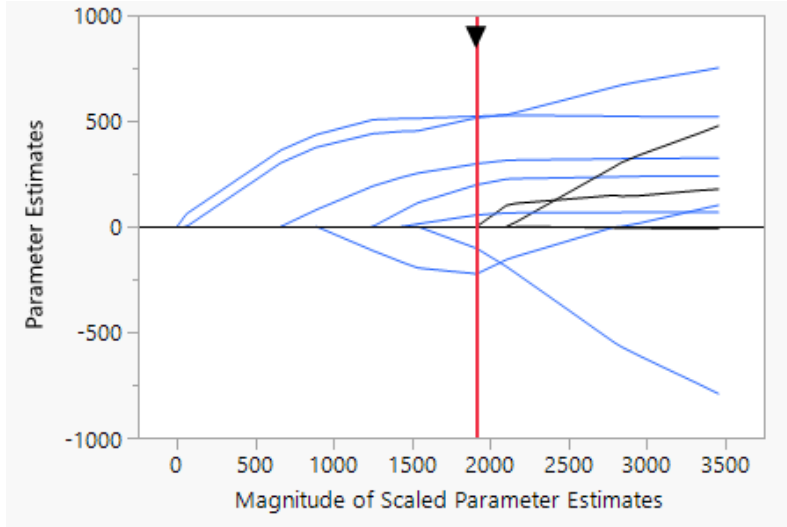
$$\hat{\beta}_{DS} = \arg \min_{\beta} \sum_j |\beta_j| \quad \text{subject to } |X^T(y - X\beta)|_{\infty} \leq s$$

In words – control the magnitude of coefficients subject to a constraint on the maximum correlation between the design and the residuals.

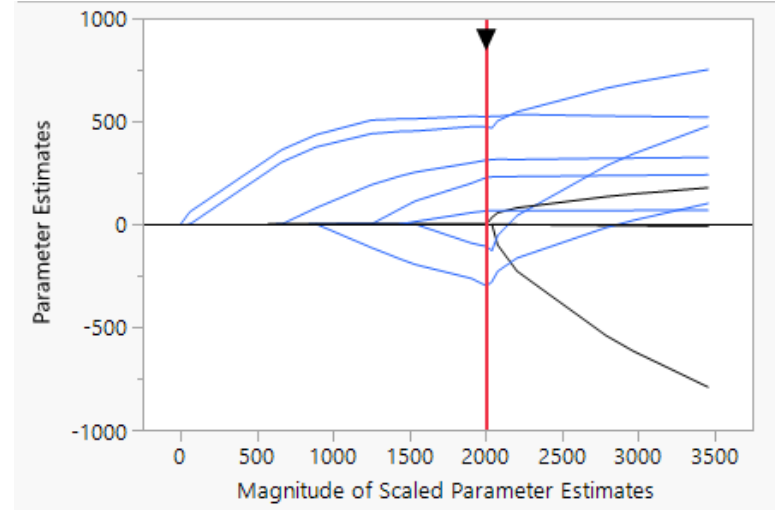
Unlike Lasso or Enet, this doesn't extend naturally to GLMs.  
Added to Genreg for 14.

# Penalized Regression

## Can you spot the difference?



Lasso



Dantzig Selector

These paths are nearly identical, but the active sets are actually slightly different.

# Penalized Regression

## Penalties and Generalized Linear Models

So far, we've focused on penalized least squares.

All of these ideas extend naturally to GLMs (except Dantzig), just penalize the likelihood rather than the sum of squared errors.

$$\hat{\beta} = \arg \min_{\beta} -\log[\textit{likelihood}(\beta)] + \lambda \sum_j \rho(\beta_j)$$

Same ideas, just fewer computational tricks.

# We talked about a lot of methods

## Rules of Thumb

Here are some rules of thumb that may help...

For designed experiments...

- Consider Forward Selection (or 2-stage version) or the Dantzig Selector.
- Stick with an information criteria for tuning.

Observational data? Consider a penalized method.

- Correlated predictors? Try the elastic net.
- ...but if all you care about is prediction, maybe the lasso.
- Use a holdback set when possible.
- Don't worry about heredity.

Don't be afraid of non-normal distributions



# Some Special Cases

# Not so obvious use cases

Our goal with Genreg is pretty ambitious:

Fit everything using a single UI

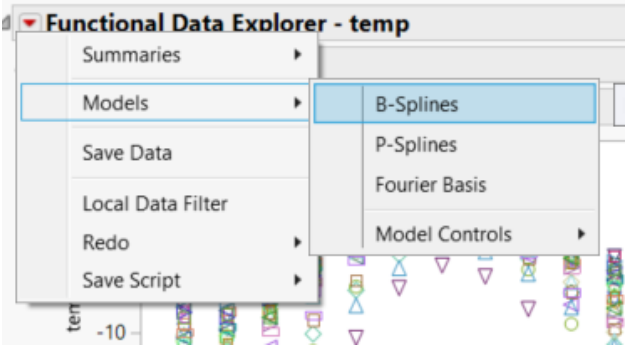
...and we've made a good bit of progress

...but now it's getting hard to hit the highlights in 90 minutes

So let's focus on a few interesting problems that Genreg can tackle.

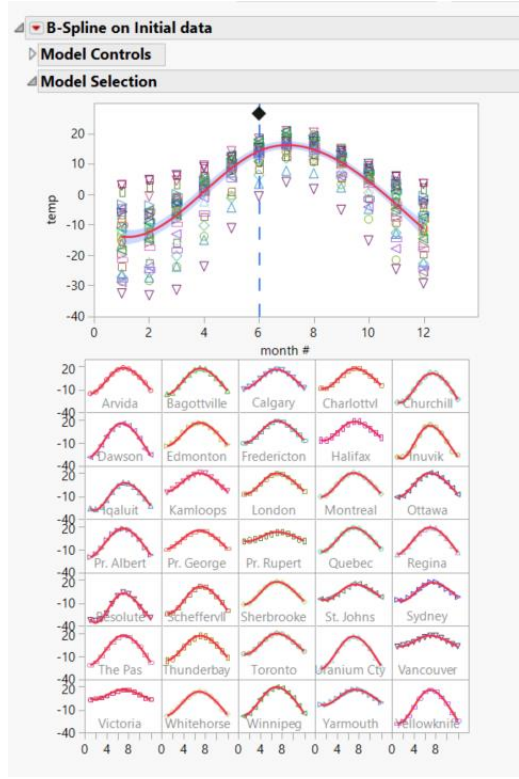
1. Functional Data
2. Censored data
3. Non-continuous responses

# Functional Data



The Functional Data Explorer was introduced in JMP Pro 14 for analyzing data that are functions, signals, series, ...

Example: temperature readings over time

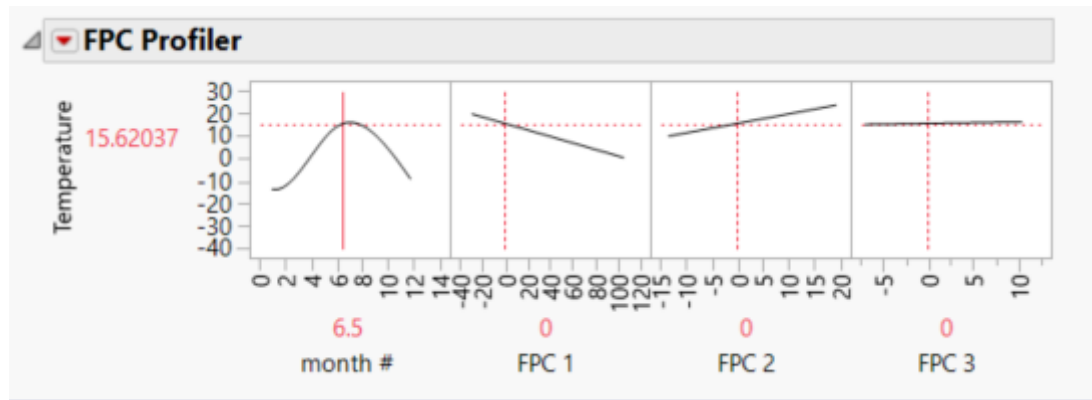


# Functional Data

Goal: understand how our response changes over time (or something else).

So we fit a spline smoother in FDE.

Now we have a model for our response as a function of time and functional principal components.



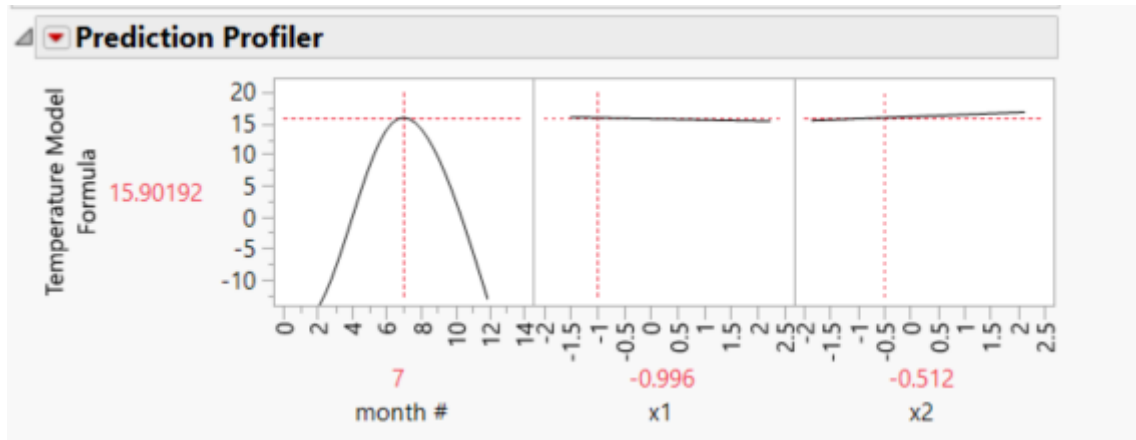


# Functional Data

But do we really care about the FPCs? Maybe not.

Instead, model the FPCs as a function of more meaningful factors  
(EX: factors from a designed experiment)

Result: model our response curve as a function of factors we care about.



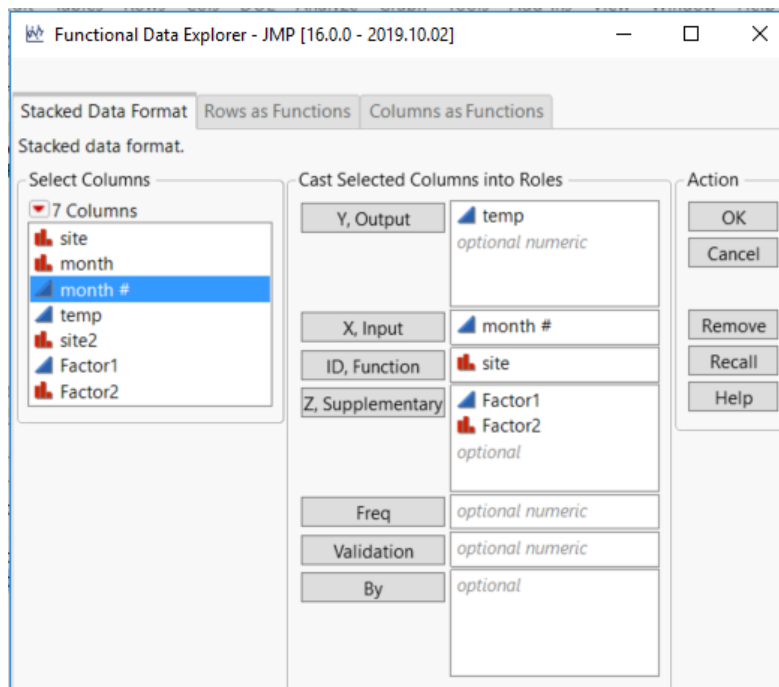
# Functional Data

## Supplementary Variables

How do we do this?

Make sure to specify the factors that we want to use as Supplementary variables.

Once we do that, we'll have two different ways to proceed...

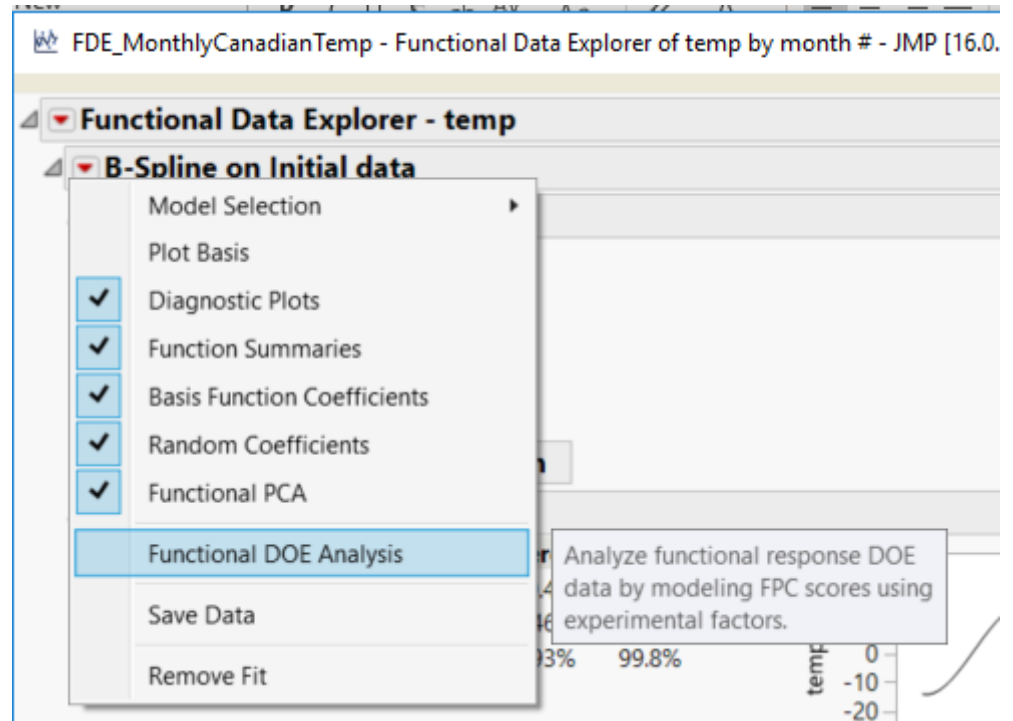


# Functional Data

## Option 1

After fitting a model in FDE, choose “Functional DOE” from the model’s red triangle menu.

This will launch Genreg within FDE and do best subset to model each functional principal component.

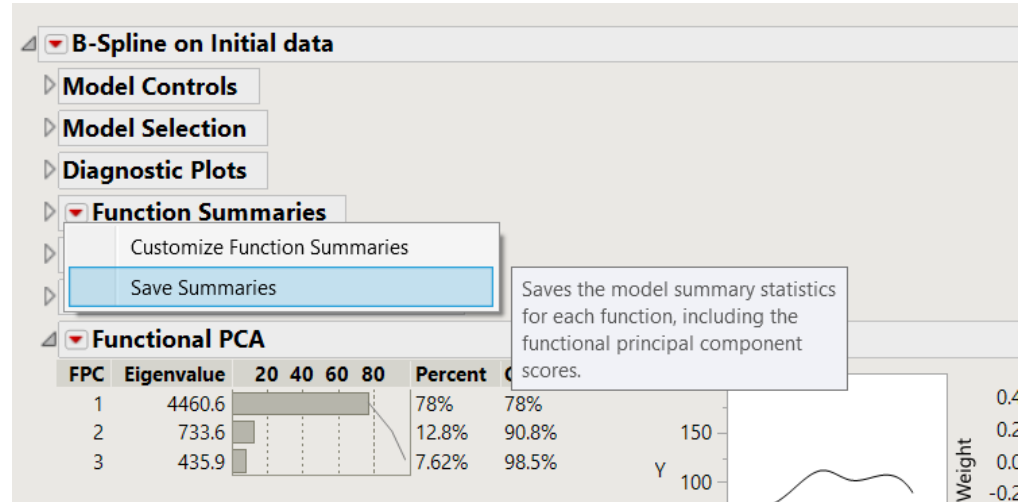


# Functional Data

## Option 2

After fitting a model in FDE, select “Save Summaries” from the Function Summaries outline.

This saves the FPCs and supplementary variables to a new data table that Genreg will recognize. This provides more control over the models you can fit.



The screenshot shows the SAS software interface for a B-Spline model. The 'Function Summaries' menu is open, showing options to 'Customize Function Summaries' and 'Save Summaries'. A tooltip explains that 'Save Summaries' saves model summary statistics for each function, including functional principal component scores.

FPC	Eigenvalue	20	40	60	80	Percent	C
1	4460.6					78%	78%
2	733.6					12.8%	90.8%
3	435.9					7.62%	98.5%

Additional visible elements include a plot of a function with 'Y' on the vertical axis (ranging from 100 to 150) and 'Weight' on the horizontal axis (ranging from -0.2 to 0.4). A small line graph is also visible in the bottom right corner.

# Functional Data

DEMO!

# Censored Data

What does it mean when we say we have censoring?

If we are not always able to observe our response exactly, we only have an upper and/or lower bound, then our response is censored.

This is not an uncommon scenario...

- Clinical trials for new therapies
- Reliability studies of new products

But there are many times we only partially observe the response

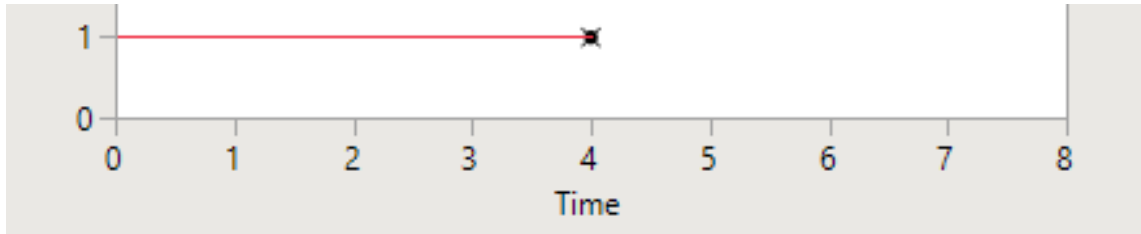
- Limit of detection (can't measure below a particular threshold)
- When the response is a sensitive subject (ex: alcoholic drinks per day)

# Types of Censoring

## No Censoring

This is the easy case – we know exactly what happened.

- We know exactly when a patient in a clinical trial dies.
- We know the exact temperature that causes our plastic to melt.

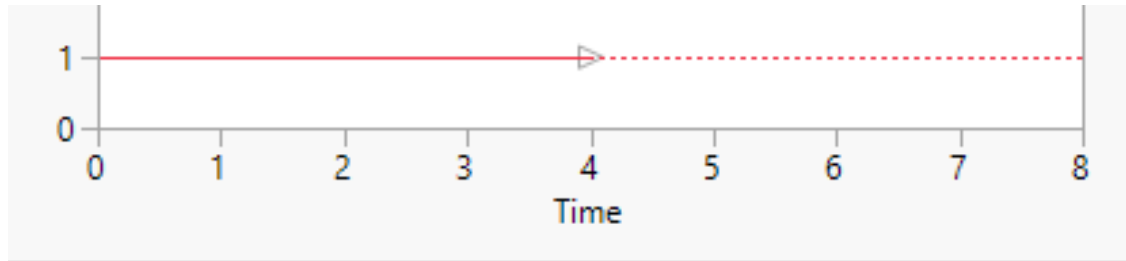


# Types of Censoring

## Right Censoring

$Y > Y_R$  – we know the lower bound on the response.

- A patient in a clinical trial is lost to follow up or the trial ends.
- Our plastic can handle the highest temperature that we can set



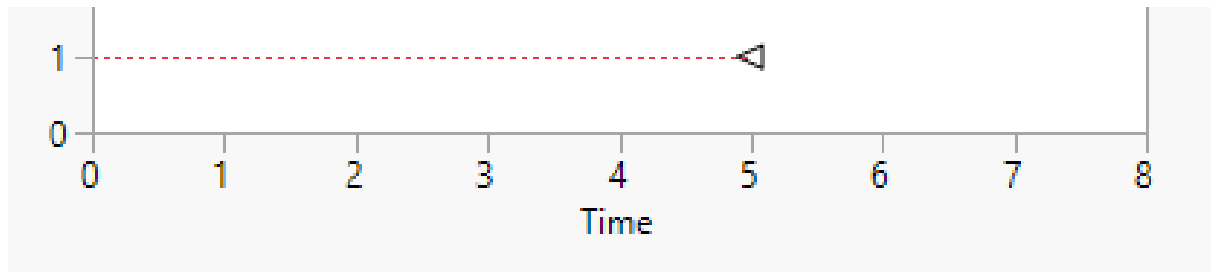


# Types of Censoring

## Left Censoring

$Y < Y_L$  – we know the upper bound on the response

- A patient enrolled in a clinical trial dies sometime before the first follow-up appointment was scheduled, but the exact time is unknown.
- A widget is smaller than  $Y_L$ , but our instruments cannot accurately measure anything below that threshold.

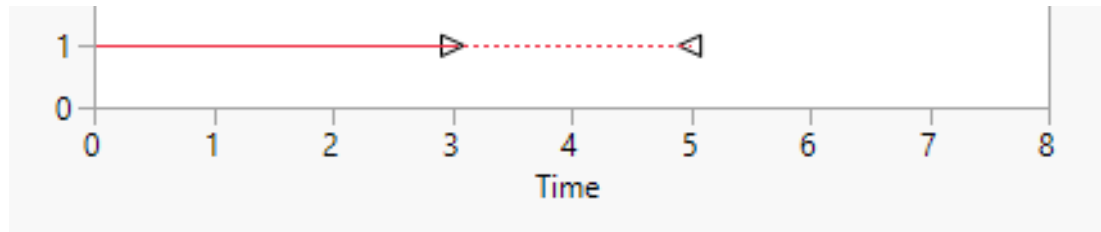


# Types of Censoring

## Interval Censoring

$Y \in (Y_L, Y_R)$  - We know the response falls between  $Y_L$  and  $Y_R$ .

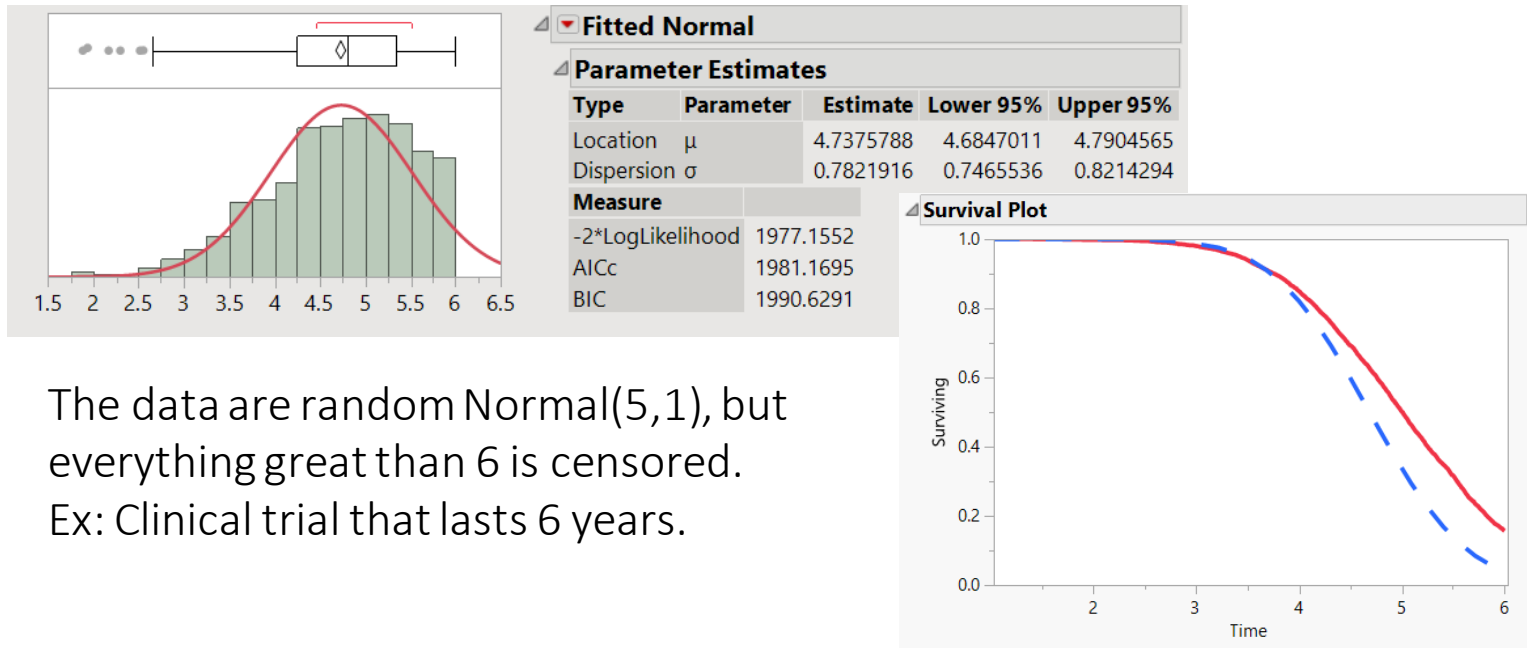
- A patient in a clinical trial dies somewhere between two follow-up appointments, but we don't know exactly when.



# What happens if we ignore censoring?

## Just drop those observations

If we drop the censored observations from our data set, we will underestimate the mean and variance.



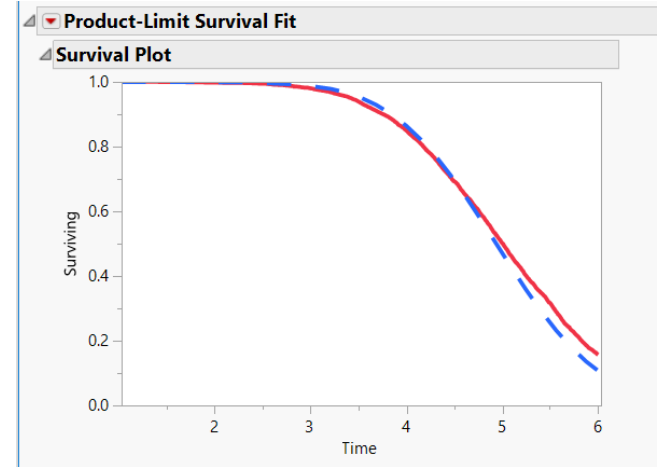
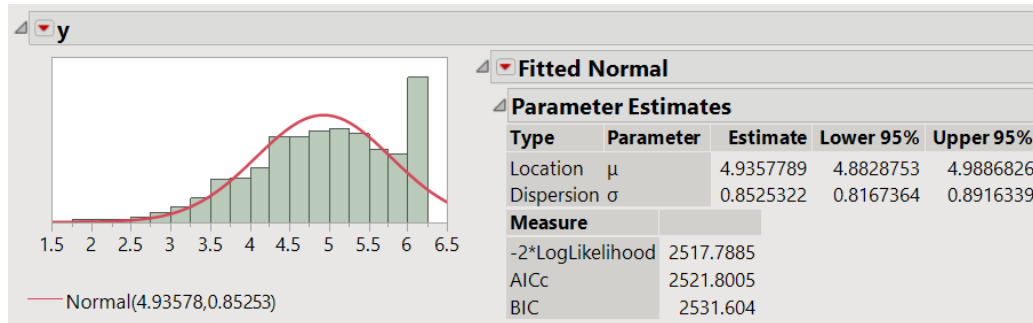
The data are random Normal(5,1), but everything great than 6 is censored.  
Ex: Clinical trial that lasts 6 years.

# What happens if we ignore censoring? Pretend the censored observations are failures

Now we end up with a big point mass at time 6.

We don't underestimate quite as badly, but still not a great result.

Neither of the CIs cover the truth.

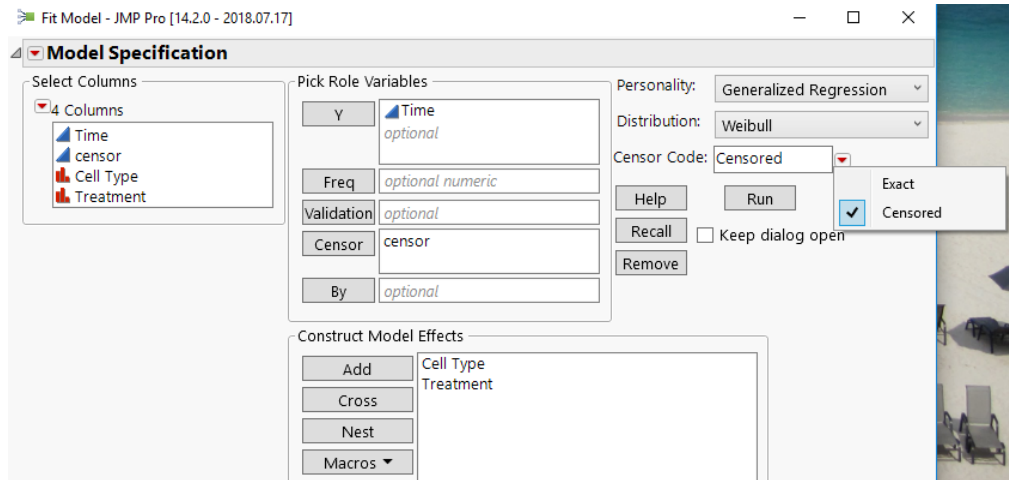




# Censored Data Format

Easiest option: designate censoring via an indicator variable

	Time	censor	Cell Type	Treatment
88	87	Exact	Small	Test
89	95	Exact	Small	Test
90	97	Censored	Small	Standard
91	99	Exact	Small	Test
92	99	Exact	Small	Test
93	103	Censored	Small	Test
94	117	Exact	Small	Standard
95	122	Exact	Small	Standard
96	123	Censored	Small	Standard
97	139	Exact	Small	Standard



But this only works for **right censoring**

# Censored Data Format

We can handle left and interval censoring, it's just a little more work...

	yLo	yHi	type	x
1	3	5	Interval	1.79
2	3	6	Interval	-1.11
3	5	5	Event	1.36
4	6	6	Event	-0.7
5	7	7	Event	0.1
6	2	2	Event	1.95
7	•	5	Left	-0.25
8	•	4	Left	0.45
9	6	•	Right	-0.65
10	7	•	Right	-0.89

We just have to provide the upper and lower endpoints on the response.

Then we specify both yLo and yHi as the response in the Genreg launch.

# Censored Data

DEMO



# Nominal and Ordinal Responses

We're used to modeling responses that take continuous values.

But that doesn't always have to be the case.

New in 14: modeling nominal (>2 levels) and ordinal responses using Genreg.

Severity makes sense to order, but probably not Medicine.

	Count	Severity	Medicine
1	151	Medium	Blue pill
2	75	Low	Blue pill
3	141	Low	Yellow pill
4	206	High	Red Pill
5	135	Low	Purple Drink
6	97	Low	Yellow pill
7	138	Low	Red Pill
8	63	Low	Purple Drink
9	110	Low	Blue pill
10	310	High	Yellow pill
11	101	Low	Blue pill
12	69	Low	Blue pill
13	179	Medium	Red Pill
14	185	Medium	Purple Drink
15	118	Low	Purple Drink
16	171	Medium	Yellow pill
17	166	Medium	Blue pill

# Ordinal Logistic Regression

## In Genreg

If our response takes values

$$l_1 < l_2 < \dots < l_m$$

Then we can write our probability model

$$\Pr(y_i = l_1) = \frac{1}{1 + \exp[-(\gamma_1 + x_i\beta)]}$$

$$\Pr(y_i \leq l_2) = \frac{1}{1 + \exp[-(\gamma_1 + \gamma_2 + x_i\beta)]}$$

$$\Pr(y_i \leq l_3) = \frac{1}{1 + \exp[-(\gamma_1 + \gamma_2 + \gamma_3 + x_i\beta)]}$$

# Ordinal Logistic Regression

Or more generally

$$\Pr(y_i \leq l_k) = \frac{1}{1 + \exp[-(\sum_{j=1}^k \gamma_j + x_i \beta)]} \quad \text{for } k < m$$

Here  $\beta$  is a vector of regression coefficients like usual.

The  $\gamma_j$  are like intercepts that separate each level of the response.

If  $x_i$  is a  $p \times 1$  vector, the entire model has  $p + m - 1$  parameters.

You may see this called the cumulative logit or ordered logit model.

# Ordinal Logistic Regression

What does our model mean in terms of odds?

Let's look at the odds ratio...

$$\frac{\text{Odds}(y = y_2)}{\text{Odds}(y = y_1)} = \frac{\exp(\gamma_1 + \gamma_2 + x\beta)}{\exp(\gamma_1 + x\beta)} = \exp(\gamma_2)$$

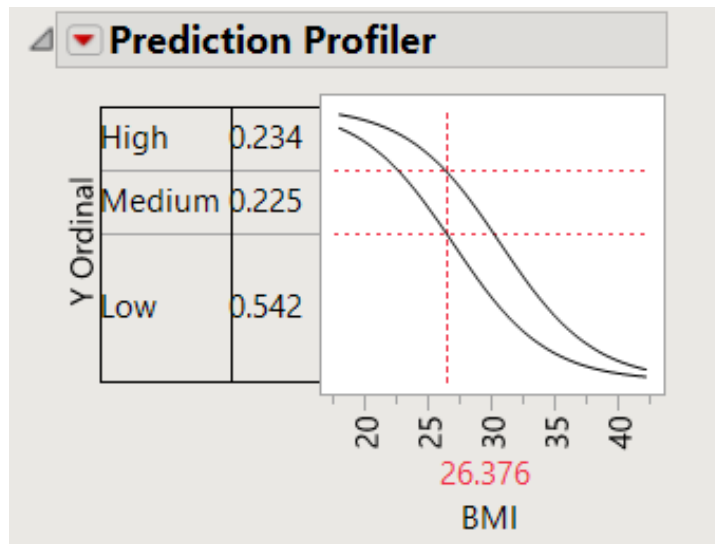
This gives us a little more intuition about these parameters.

(But does it really? I think they're pretty hard to interpret.)

You sometimes see this called the proportional odds assumption.

# Ordinal Logistic Regression

In general, the probabilities will more or less move in synchrony.



# Multinomial Distribution

If our response doesn't have a natural ordering, it gets more complicated.

Let's say our response takes values  $\{y_1, y_2, \dots, y_m\}$

Ex: { hot dog, pizza, hamburger } or { blue, green, purple, red }

$$\Pr(Y_i = y_j) = \frac{\exp(x_i \beta_j)}{1 + \sum_{k=1}^{m-1} \exp(x_i \beta_k)} \text{ for } j < m$$

$$\Pr(Y_i = y_m) = 1 - \sum_{j=1}^{m-1} \Pr(Y_i = y_j)$$

# Multinomial Distribution

So each level of the response (except the last) gets its own set of regression parameters  $\hat{\beta}_j$ .

Let's say we have  $p = 15$  predictors that we want to include in our model and our response has  $m = 5$  levels.

$$\begin{aligned}\text{Number of parameters} &= (p + 1) * (m - 1) = 16 * 4 \\ &= 64\end{aligned}$$

For comparison, if we could order the response we'd have 19 parameters.

# Multinomial Distribution

**Model Specification**

Select Columns: 7 Columns  
Color, x1, x2, x3, x4, x5, Validation

Pick Role Variables:  
Y: Color  
Weight: optional numeric  
Freq: optional numeric  
Validation: Validation  
By: optional

Construct Model Effects:  
Add: x1, x2, x3, x4, x5  
Cross  
Nest  
Macros

Personality: Generalized Regression  
Distribution: Multinomial  
Help, Run, Recall, Remove, Keep dialog open

**Parameter Estimates for Original Predictors**

Term	Estimate
Blue:Intercept	-16.86322
Blue:x1	0.0039636
Blue:x2[1-2]	0.215846
Blue:x3	0.3094224
Blue:x4	0.0651246
Blue:x5	0.0059876
Brown:Intercept	-7.190225
Brown:x1	0.0044349
Brown:x2[1-2]	0.46059
Brown:x3	0.112067
Brown:x4	0.0329118
Brown:x5	-0.001602

We add the name of the response level for each parameter in the model.

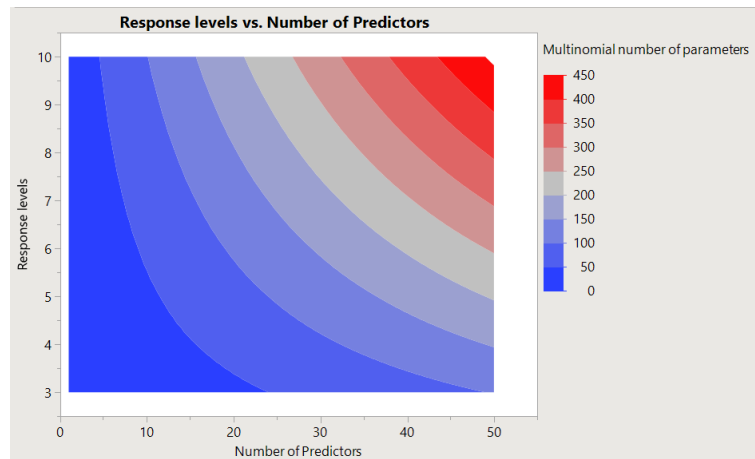
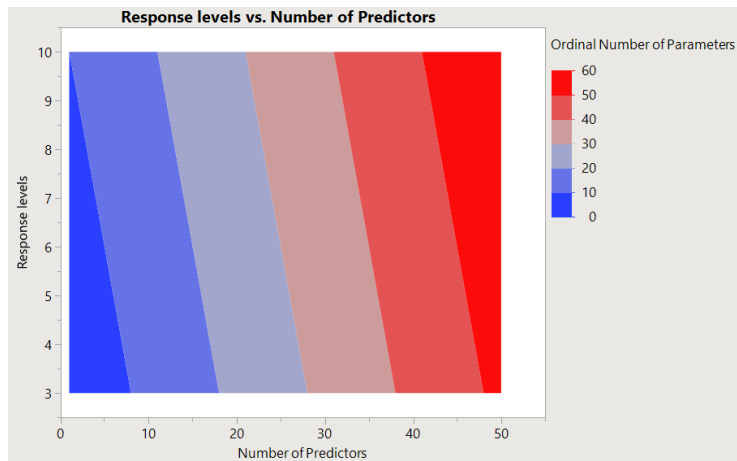


# Ordinal and Multinomial

Recall that with  $p$  predictors and an  $m$  level response,

Ordinal fits  $p+m-1$  parameters

Multinomial fits  $(p+1)*(m-1)$  parameters



# Ordinal and Multinomial

- We could try to order the response to keep the model manageable.
- But with the penalized regression tools in Genreg, we don't have to worry as much about overparameterizing the multinomial.
- In fact, should we try multinomial even when the response is ordered?

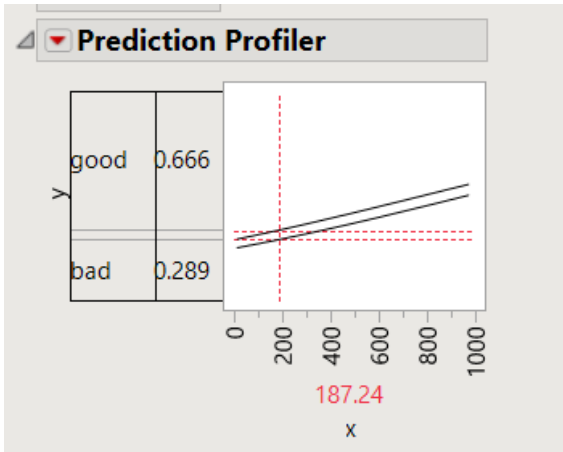
Measures of Fit for 4-Level Nominal Response						
Validation	Creator	.2 .4 .6 .8	Generalized RSquare	Misclassification Rate	N	
Training	Fit Generalized Lasso		0.5063	0.2945	309	
Training	Fit Generalized Maximum Likelihood		0.7350	0.2039	309	
Validation	Fit Generalized Lasso		0.3554	0.3008	133	
Validation	Fit Generalized Maximum Likelihood		-2.388	0.5038	133	

# Ordinal and Multinomial

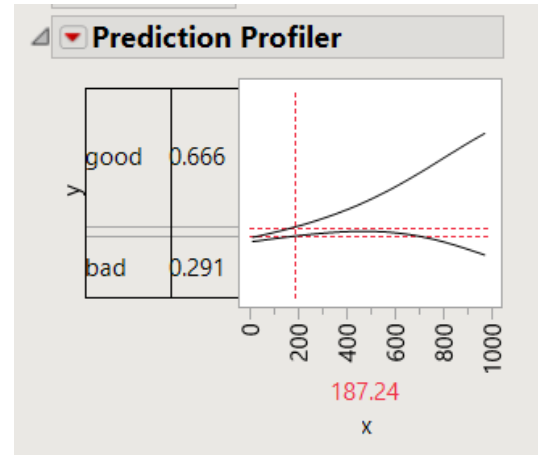
Earlier we saw that the ordinal probabilities are sorta parallel.

The extra parameters in the multinomial give us much more flexibility.

Ordinal



Multinomial



# Ordinal and Multinomial

DEMO



# Wrap-up

# Wrap-up

## What We've Learned

Genreg provides a variety of variable selection methods.

Genreg is efficient – a single UI means that learning a little bit of Genreg will help you solve a lot of different problems.

New in 15: easier comparison of models.

Genreg can be your one-stop model building platform in JMP Pro.

# Wrap-Up

## Related talks

Thursday 1:45-2:30

### Not Quite Normal: Choosing the Best Distribution for Modeling Your Response

Clay Barker, JMP Principal Research Statistician Developer, SAS

Friday 9:00-9:45

### Analysis of Fly Ash Concrete Curing Curves From a Mixture Amount Experiment Using FDE in JMP<sup>®</sup> Pro

Philip Ramsey, Principal Lecturer, University of New Hampshire; and Owner, North Haven Group

Christopher Gotwalt, JMP Director of Statistical Research and Development, SAS

# Wrap-Up

## Related Talks

### Anomaly Detection and JMP<sup>®</sup> Pro

Michael Crotty, JMP Senior Statistical Writer, SAS

Colleen McKendry, JMP Technical Writer, SAS

Marie Gaudard, Statistical Consultant

Thursday  
11:00-11:45

### Text Curation Example Using Genreg Platform

Scott Reese, Senior Scientist, Procter & Gamble

Amy Phillips, Principal Scientist, Procter & Gamble

Tracy Desch, Scientist, Procter & Gamble

A. Narayanan, Principal Scientist, Procter & Gamble

Thursday  
1:00-1:30



# Wrap-up

## More Resources

Search the JMP blog for posts about Genreg.

---

**JMP 13 Preview: More enhancements to generalized r...**

09-14-2016 01:17 PM 

by  [anne\\_milley](#) in [JMP Blog](#)

1 Kudo 

Clay Barker has been busy extending the usefulness of the Generalized Regression platform in JMP Pro, adding many new models and enhancing ease of use. Generalized Regression (or GenReg for short) de...

Tags:  [generalized regression](#)  [jmp 13](#)  [jmp pro](#)  [statistics](#)

Search our user community (<https://community.jmp.com/welcome>) for old presentations (by myself and others).

---

**Visually Exploring Design of Experiments Models Wi...**

08-04-2016 03:56 PM 

by  [kathy\\_walker](#) in [Discovery Summit 2016 Presentations](#)

Chris Gotwalt [chris.gotwalt1](#), PhD, JMP Director for Statistical Research and Development, SAS Clay Barker [clay.barker](#), PhD, JMP Senior Research Statistician, SAS The Generalized Regression p...



Thanks!  
Clay.Barker@sas.com

[sas.com](https://sas.com)