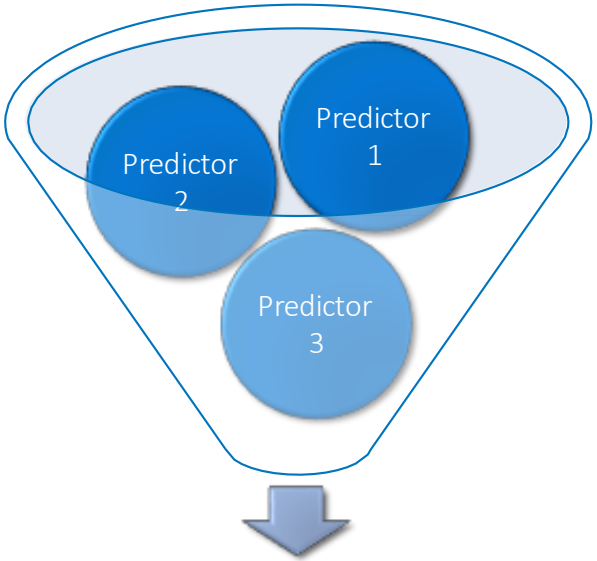
Generalized Regression

(Clay Barker and Chris Gotwalt, JMP)

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!
- Generalized Regression helps!

What is Variable Selection?

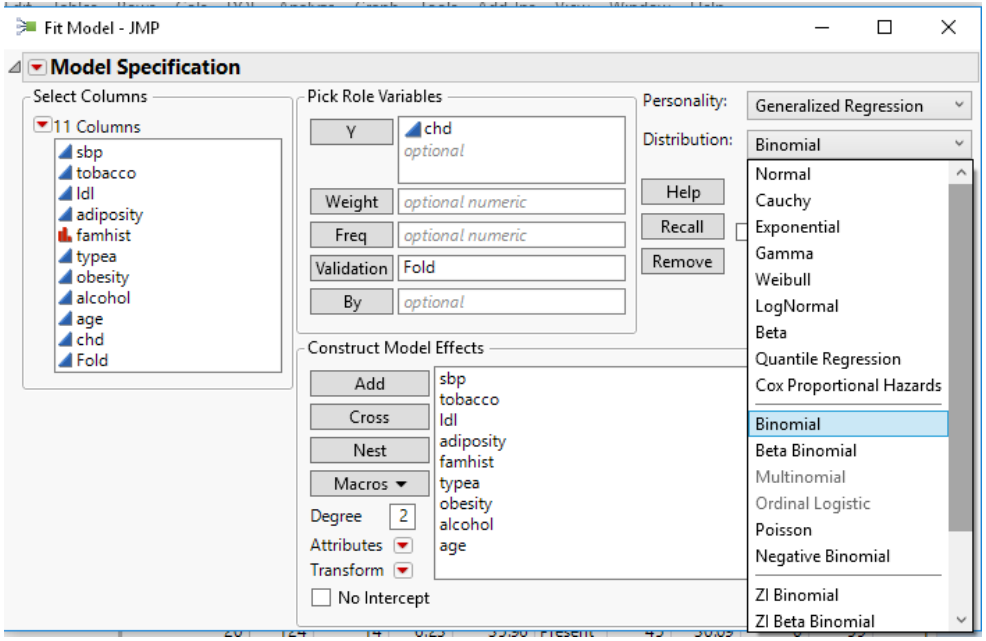
- Variable selection is crucial for several reasons.
- The resulting model…
  1. …is easier to interpret. Often it is much easier to interpret
  2. …will generalize well to new observations.
  3. …is stable to small changes in the observed data.
  4. …is easier to use/deploy.

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

The Generalized Regression Platform

What is it?

- Fit Model personality introduced in JMP Pro 11 called Genreg.



Genreg

Response Distributions

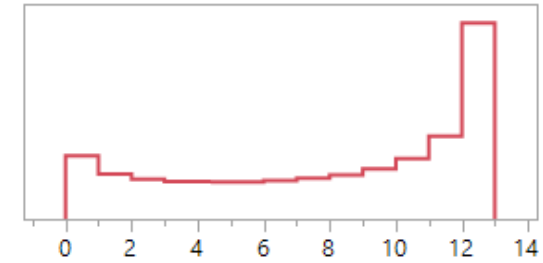
- Genreg can handle a wide variety of response types. We can't always assume that our response is normally distributed.

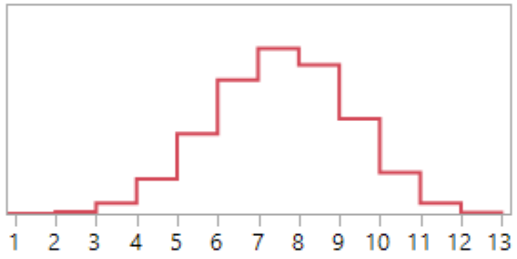
- Skewed – Gamma*, Weibull*, Lognormal*, Exponential*, Beta
- Count – Negative Binomial, Poisson, Binomial, and zero-inflated versions.
- Label – Binomial, Multinomial, and Ordinal Logistic
- Other – Quantile Regression, Cauchy, Proportional Hazards*

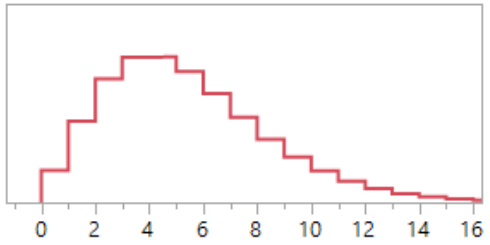
* supports censoring

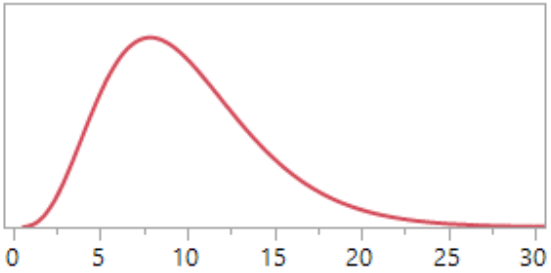
Genreg

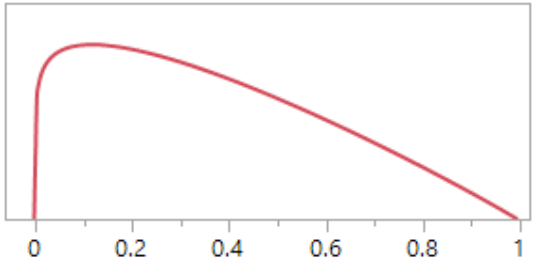
This covers a lot of cases

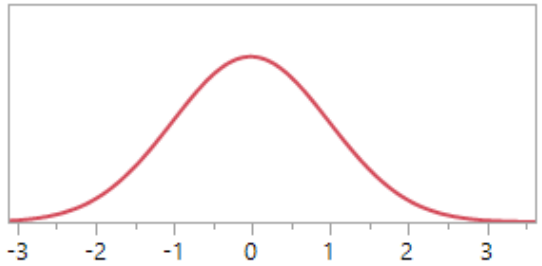












Genreg

Estimation and Selection

- Genreg has a variety of estimation 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

- Genreg's goal is to provide a single unified framework for interactively building models in a wide variety of settings.
  - …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.

Stepwise Methods in Genreg

- Genreg offers a handful of *stepwise* variable selection methods.
- Why do we call them stepwise?
  - These are largely algorithmic methods.
  - Given our 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 and keep the best.

- 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.

Seems reasonable, right?

Stepwise Methods in Genreg

Best subset

- Not always feasible! Consider 10 main effects, 10 quadratics, and all possible 2 factor interactions (45 of them)

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

Stepwise Methods in Genreg

Forward Selection

- Best Subset is not feasible for even moderately sized problems.
- Instead of fitting literally everything, Forward Selection uses heuristics to help us choose a competitive model of each model size.
- 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

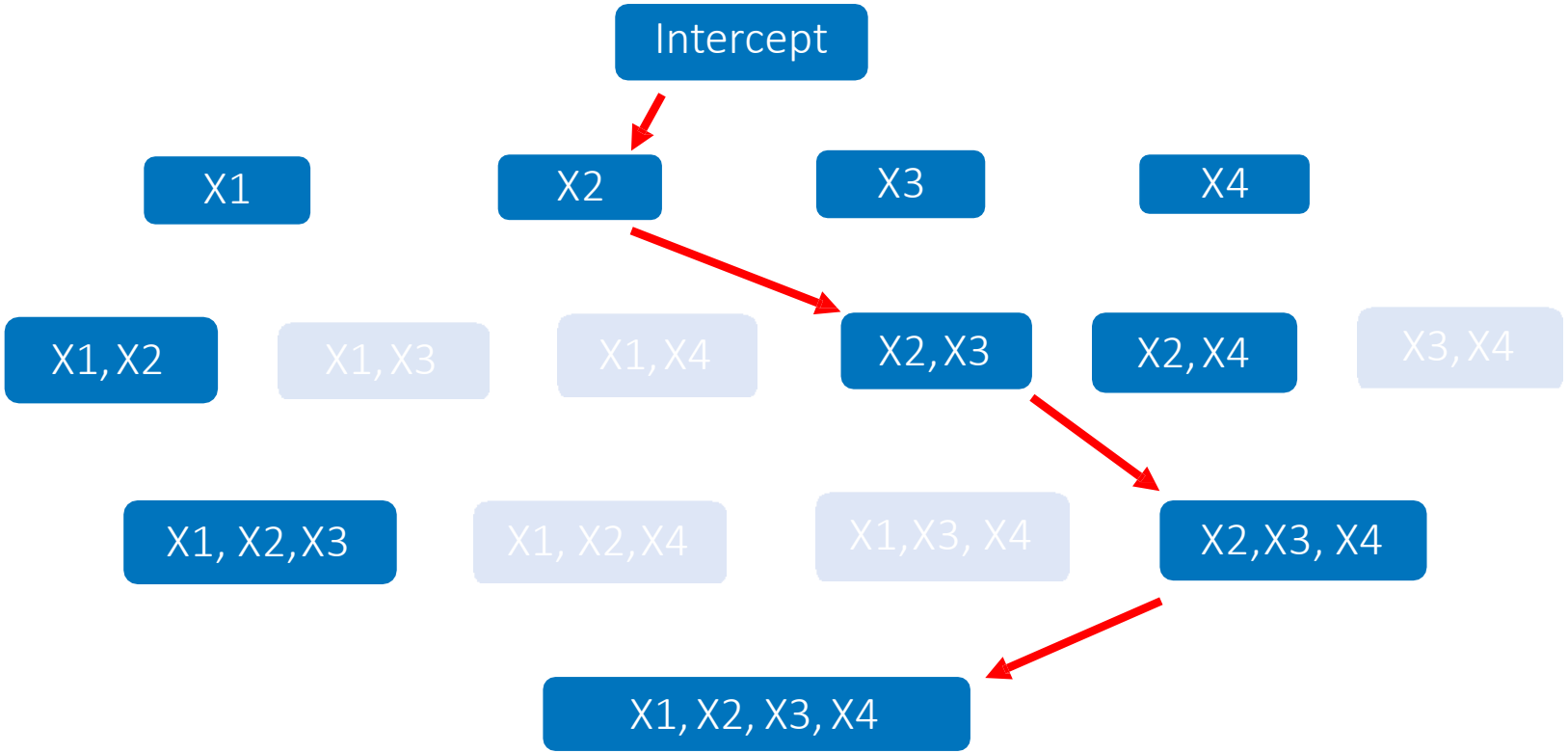
- Consider a (very) simple example with 4 predictors.

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

- Now we have 5 models to consider out of 16 possible
  1. Interceptonly
  2. X2
  3. X2, X3
  4. X2, X3, X4
  5. X2, X3, X4, X1

Stepwise Methods in Genreg

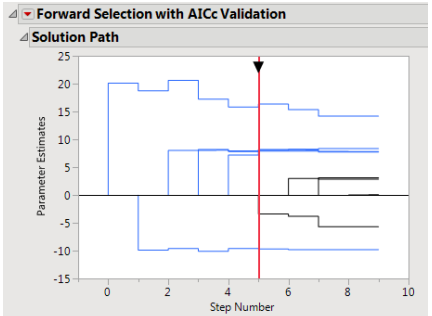
Forward Selection



Stepwise Methods in Genreg

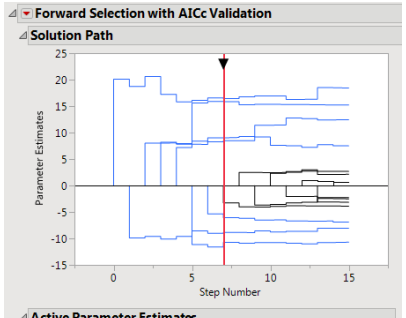
Two-Stage Forward Selection

- 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.



Active Parameter Estimates

Term	Estimate	Std Error	Wald ChiSquare
Intercept	-5.538269	0.8809954	39.518531
tobacco	0.0803753	0.0250518	10.293569
ldl	0.1619916	0.0559878	8.3714176
famhist[Absent-Present]	-0.908175	0.2248549	16.313035
typea	0.0371152	0.0118111	9.8747155
age	0.0504604	0.00966	27.286505



Active Parameter Estimates

Term	Estimate	Std Error	Wald ChiSquare
Intercept	-6.970279	0.9999311	48.591483
tobacco	0.0837833	0.0257068	10.622289
ldl	0.3572134	0.089388	15.96972
famhist[Absent-Present]	-0.851191	0.2322616	13.430733
typea	0.0427935	0.0122257	12.252042
age	0.0527964	0.0099233	28.307085
(tobacco-3.63565)*(typea-53.1039)	-0.005308	0.0023366	5.1609842
(ldl-4.74032)*famhist[Absent-Present]	-0.346879	0.1137811	9.2942844

Stepwise Methods in Genreg

Two-stage Forward Selection

- Why break the selection process up?

- Especially nice for analyzing designed experiments
  1. We believe main effects are stronger than higher order effects.
  2. We may be more aggressive/liberal in the first stage.
  3. We believe in effect heredity.

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.

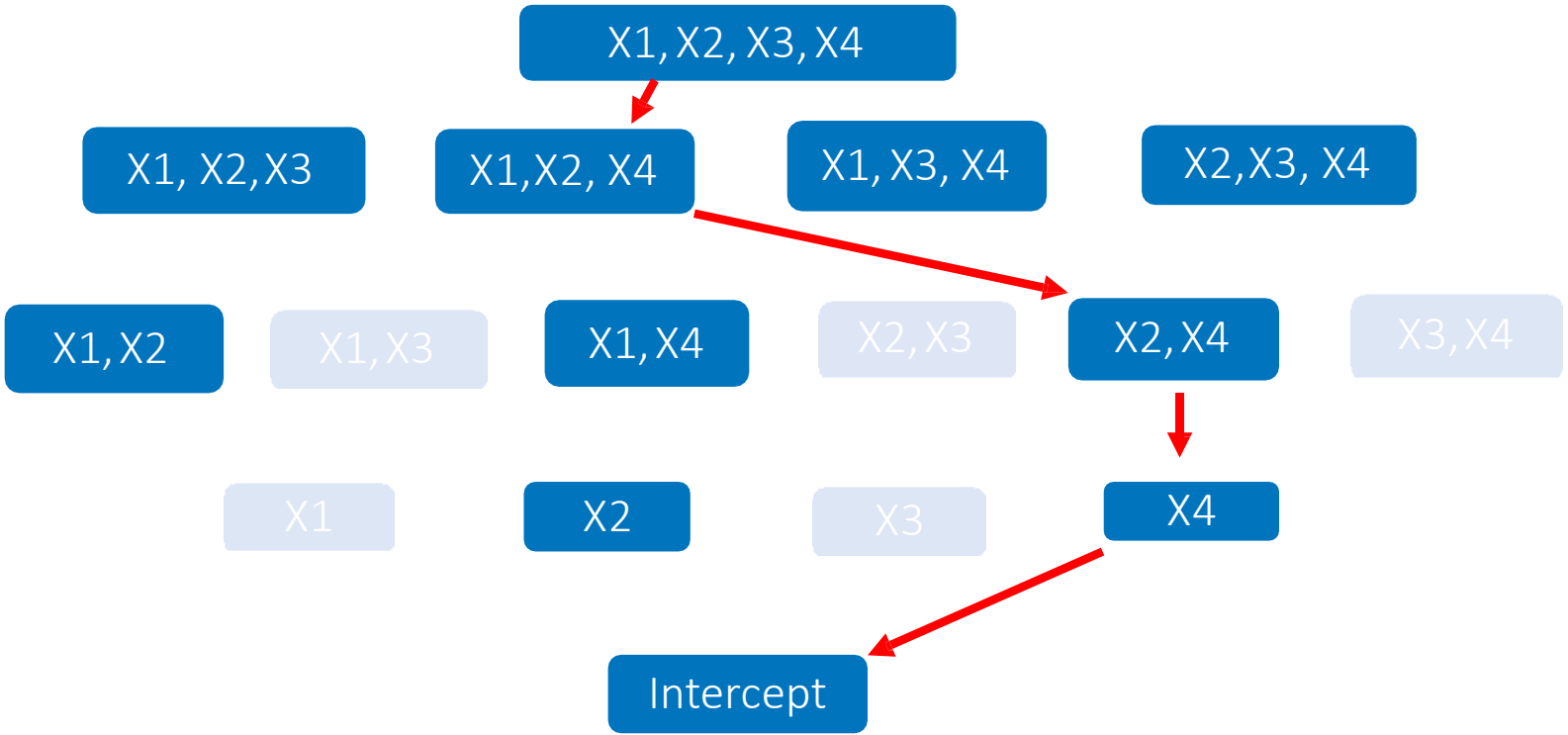
	Step 1	Step 2	Step 3	Step 4
p for X1	.001	.4*		
p for X2	.2	.15	.06*	
p for X3	.6*			
p for X4	.05	.03	.01	.003

- Again we have a sequence of 5 models to 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



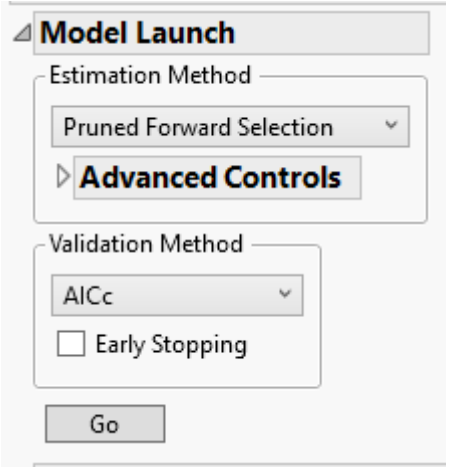
Stepwise Methods in Genreg

Forward and Backward Steps

- There are good things about both Forward and Backward Selection-

  Wouldn't it make sense to combine them?

- That's our goal with the Pruned Forward Selection method in Genreg.
  - Unique to Genreg
  - At each step in the algorithm, we consider adding a term, dropping a term, or swapping them.



Stepwise Methods in Genreg

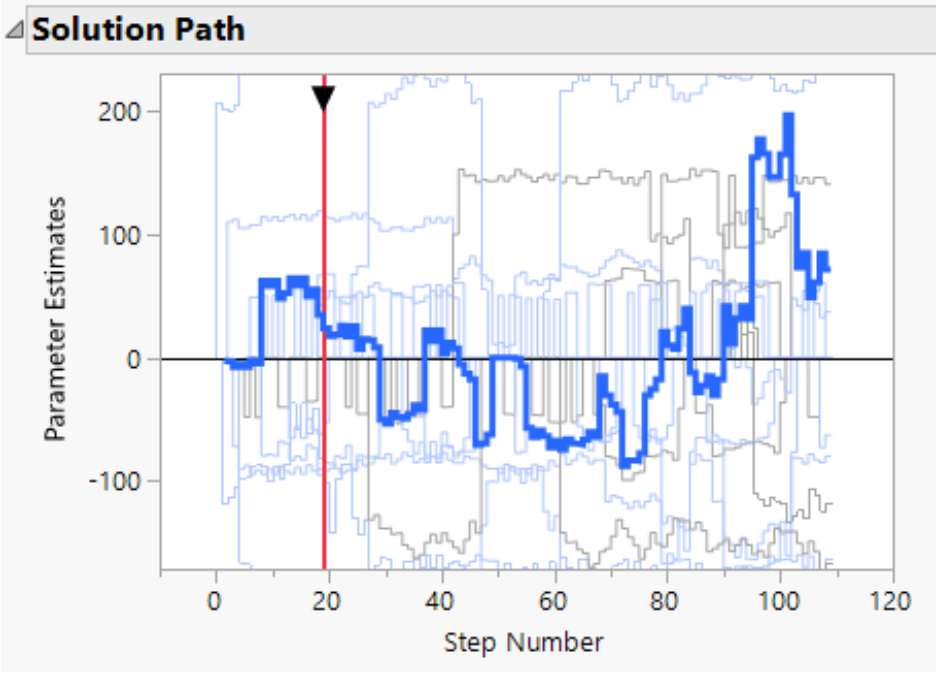
Pruned Forward Selection

- 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.
- Be careful not to get stuck in a loop.

Stepwise Methods in Genreg

Pruned Forward Selection

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



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

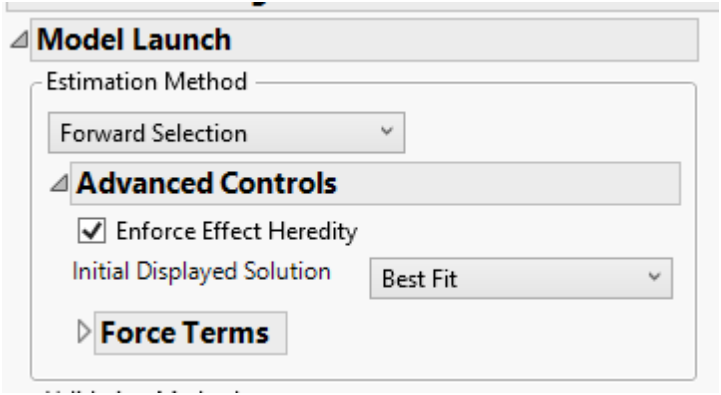
- This is sometimes called strong effect heredity.

Stepwise Methods in Genreg

Effect Heredity

- Effect Heredity option is in the Advanced Controls panel.
- Heredity often makes sense for designed experiments.

  If we know your table is a designed experiment, we enforce heredity by default.



- Heredity is quite polarizing – some people love it, others don't.

Demonstration

Model Selection with MLE and Stepwise Methods

Penalized Regression

- Stepwise methods are great.
  1. Easy to implement
  2. Intuitive and easy to explain

- But stepwise methods aren't the only school of thought.

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

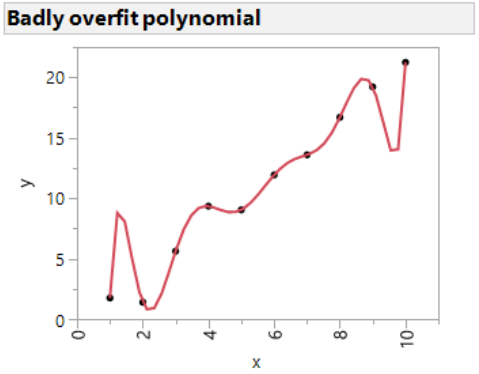
Penalized Regression

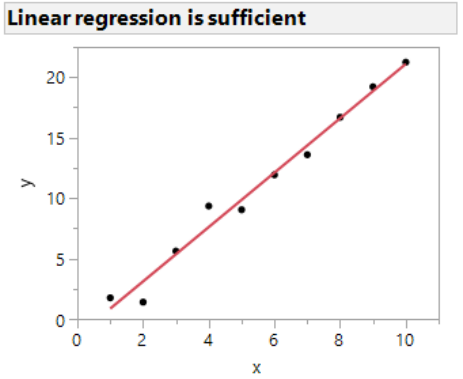
Overfitting

- 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.

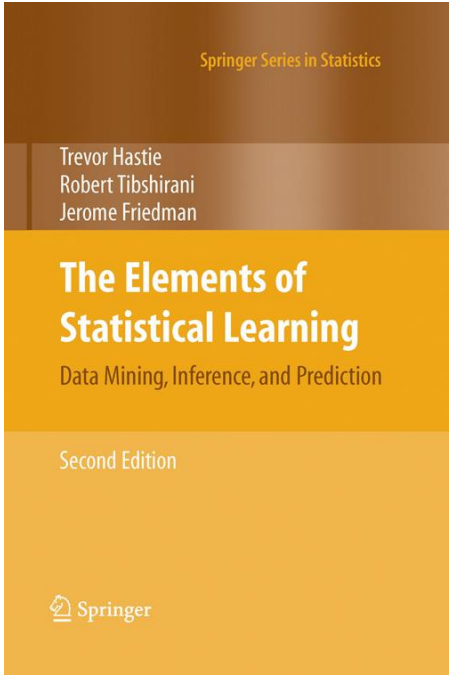
- Classic overfitting
  - Our model fits great on the observed data ☺
  - Our model fails miserably when predicting new observations
  - Our inferences are misleading
  - If we slightly alter the data, our model changes dramatically



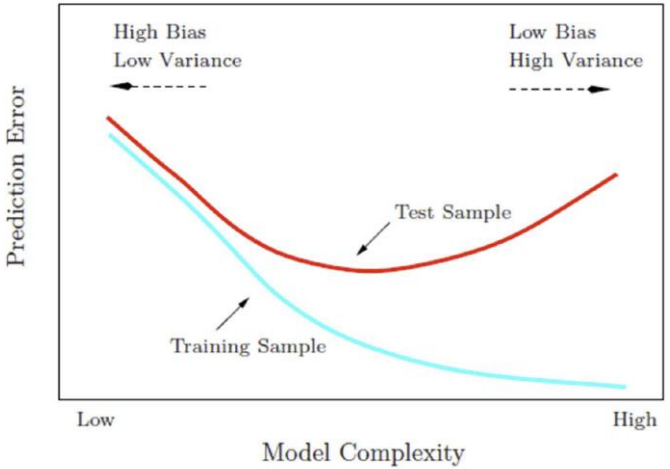


The Ubiquitous Bias-Variance Balance

An Excellent Reference



This provides a detailed, mathematically rigorous approach to data mining. It is available for free from http://www-stat.stanford.edu/~tibs/ElemStatLearn/download.html



Penalized Regression

Prediction Error

- Before we get to penalized methods, let's talk about prediction error.
- Suppose we observe data of the form

$$Y_i = f(X_i) + \epsilon_i \quad i = 1, \ldots, n$$

$$\epsilon \sim N(0, \sigma^2) \qquad X_i \text{ is } p \times 1 \text{ vector of predictors}$$

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

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

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

Penalized Regression

Prediction Error

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

$$= \mathrm{E}\{\left[y_{n+1} - f(x_{n+1}) + f(x_{n+1}) - \hat{f}(x_{n+1})\right]^2\}$$

$$= E\{[y_{n+1} - f(x_{n+1})]^2\} + E\left\{\left[f(x_{n+1}) - \hat{f}(x_{n+1})\right]^2\right\}$$

$$+ 2E\{[y_{n+1} - f(x_{n+1})]\left[f(x_{n+1}) - \hat{f}(x_{n+1})\right]\}$$

We know that $E[y_{n+1} - f(x_{n+1})] = 0$ because $\mathrm{E}\ (\epsilon_i) = 0$.

$$\Rightarrow\ PE\left(\hat{f}(x_{n+1})\right) = \sigma^2 + \underbrace{E\{\left[f(x_{n+1}) - \hat{f}(x_{n+1})\right]^2\}}_{\text{MSE}}$$

$$PE\left(\hat{f}(x_{n+1})\right) = \sigma^2 + MSE\left(\hat{f}(x_{n+1})\right)$$

Penalized Regression

The Bias/Variance Tradeoff

- Prediction Error $\left(\hat{f}(X_{n+1})\right) = \sigma^2 + \text{MSE}\left(\hat{f}(X_{n+1})\right)$

$$= \sigma^2 + \text{E}\left[f(X_{n+1}) - \hat{f}(X_{n+1})\right]^2 + \text{var}\left[\hat{f}(X_{n+1})\right]$$

Fixed — Bias Squared — Variance

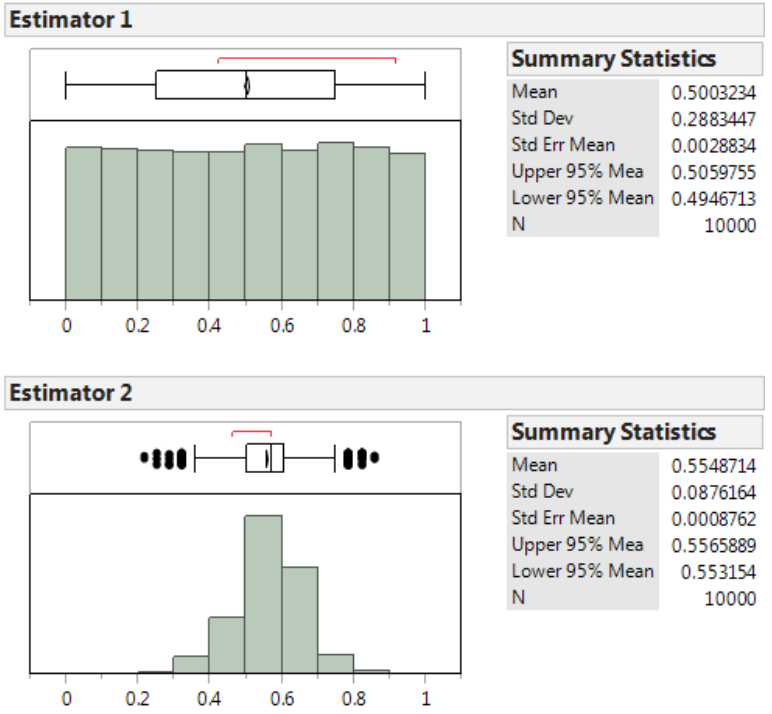
- This is the bias/variance tradeoff in estimation.
- Stepwise methods use maximum likelihood estimation, which is unbiased.
- Maybe we can accept some bias to reduce variance?

  This is the motivation behind penalized regression!

Penalized Regression

An exaggerated example of bias/variance tradeoff

- This tradeoff comes up all the time in statistics.
- The estimator on the top is unbiased but highly variable.
- The estimator on the bottom is biased, but much less variable.



Penalized Regression

Ridge Regression

- OLS is unbiased and we worked out the prediction error.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i} (y_i - x_i\beta)^2 = (X^T X)^{-1} X^T y$$

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

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \sum_{i} (y_i - x_i\beta)^2 + \frac{\lambda}{2} \sum_{j} \beta_j^2$$
$$= \left(X^T X + \lambda I_p\right)^{-1} X^T y$$

- Tuning parameter λ controls the magnitude of parameters.
  - $\lambda = 0$ is the usual OLS solution
  - As λ increases, parameter estimates move toward zero. Shrinkage!

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.

Demonstration

Ridge Regression Diabetes Data

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)
$\lvert x \rvert$	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

- There are 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:

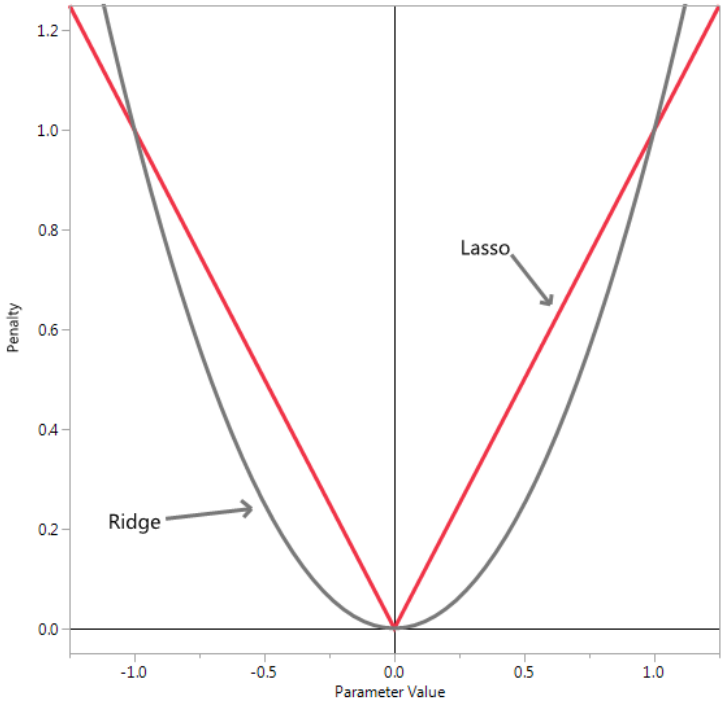
$$\hat{\beta}_{lasso} = \arg\min_{\beta} \sum_{i} (y_i - x_i\beta)^2 + \lambda \sum_{j} |\beta_j|$$

- 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 difficult (derivative undefined at zero).
- Better predictions and interpretation are worth the extra trouble!

Penalized Regression

Ridge vs Lasso penalty



- Close to zero, the lasso penalty is much more harsh than the ridge penalty.
- This partially explains why lasso is able to shrink parameters all the way to zero but ridge cannot.

Penalized Regression

Back to Diabetes

- Unfortunately the lasso MSE is not easy to work out.
  - As λ increases, bias goes up and variance comes down.
  - 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

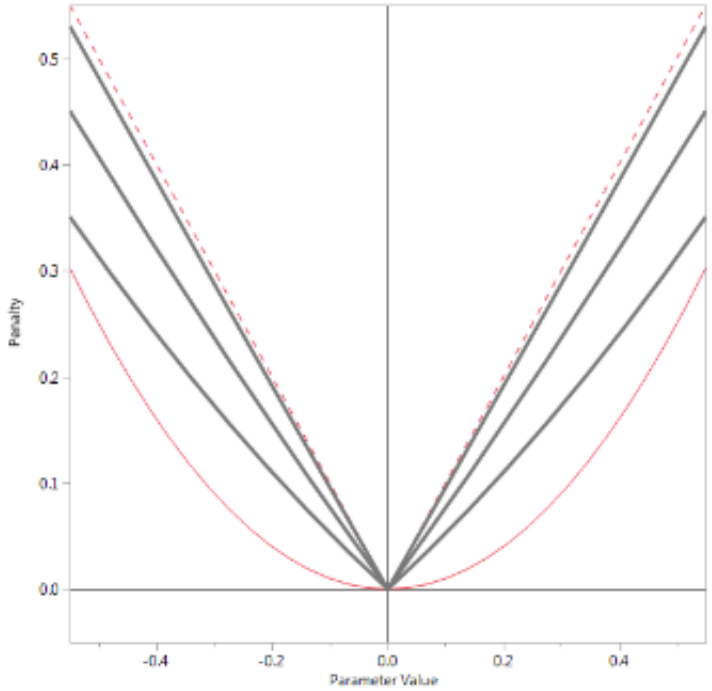
- Zou and Hastie (2005): Ridge + Lasso = Elastic Net

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

- α tuning parameter controls the mix of ℓ_1 and ℓ_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 α close to 1 (default is .99 in Genreg)

Penalized Regression

Elastic Net Penalty



Penalized Regression

Elastic Net vs Lasso

- Suppose we have 10 candidate predictors.
- x_2 and x_4 are highly correlated and they are both truly active.
  - Lasso will likely only choose x_2 or x_4
  - Elastic Net will likely choose x_2 and x_4
- So which solution is better? It probably depends on context.
- The other difference? Elastic net can select more than *n* parameters, which may or may not be a good thing.

Penalized Regression

Diabetes

- 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

Adaptive Lasso

- What if we knew in advance which predictors are important?

  Then variable selection seems unnecessary…

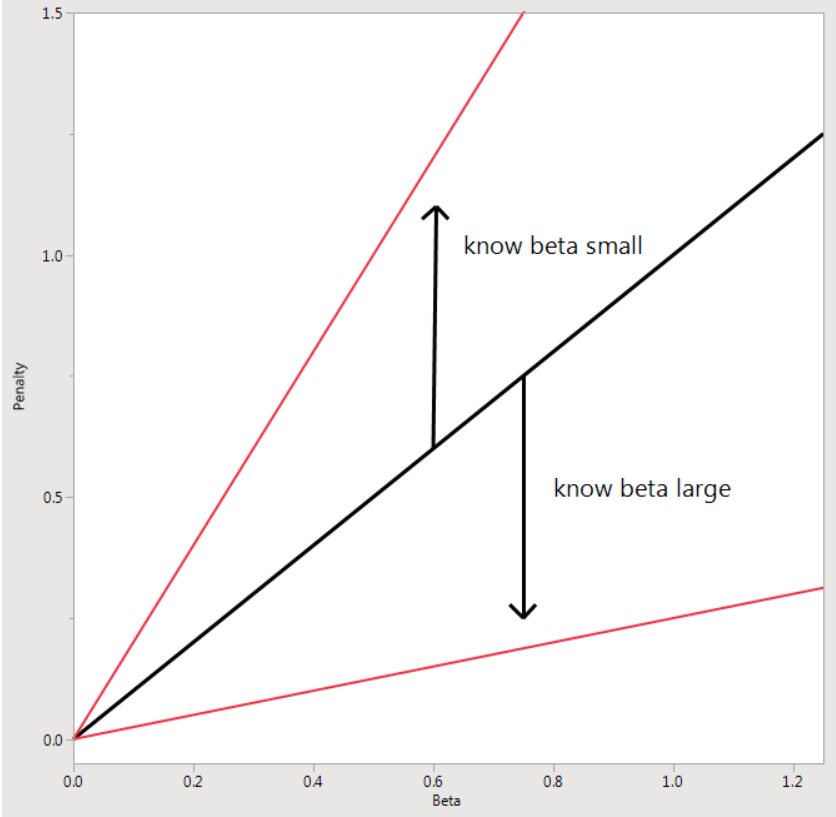
- But regardless if we somehow knew which predictors were important, we might penalize their coefficients less.

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

- A predictor that we know is important would get a smaller weight.

Penalized Regression

Adaptive Lasso



Penalized Regression

Adaptive Lasso

- Carefully chosen weights give the adaptive lasso the *oracle property*. That means that asymptotically,
  1. We should choose the correct active set.
  2. We should predict as well as if we knew the true active set in advance.

- If we use the inverse of the OLS solution, we get the oracle property.

$$w_j = {}^{1}/_{|\hat{\beta}_{j,OLS}|}$$

Penalized Regression

Adaptive Lasso

- If OLS estimates are unstable, the adaptive lasso may be poor
- The nice theory around the adaptive lasso may be based on assumptions that are not appropriate for your data.

- You may want to avoid the adaptive lasso when
  1. You have singularities ($n \ll p$)
  2. Your predictors are highly correlated
  3. Your adaptive lasso fit looks suspicious

- Bottom line is to be careful with this one

Penalized Regression

Another variation of the Lasso

- There could be a benefit to doing the lasso twice.
  1. Do the lasso on the full set of predictors, giving us a set S.
  2. Do the lasso on S.

- This is called the Double Lasso. Why do two passes?

  Pass 1 = Selection

  Pass 2 = Shrinkage

- Breaking the process in two parts helps avoid *overshrinking*, which can result in a better model.

Penalized Regression

Penalties and Generalized Linear Models

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

- All of these ideas extend naturally to GLMs, just penalize the likelihood.

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

- Same ideas, just fewer computational tricks.

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.

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

  - In words – control the magnitude of coefficients subject to a constraint on the maximum correlation between the design and the residuals.
  - Unlike Lasso or Elastic Net, this doesn't extend naturally to GLMs.

Penalized Regression

The Dantzig Selector

- The form of the Dantzig Selector is quite unique, but it has nice theoretical properties and has shown promise in analyzing designed experiments.

- Solution is very similar to the Lasso in many cases
  - Consider for supersaturated designs.

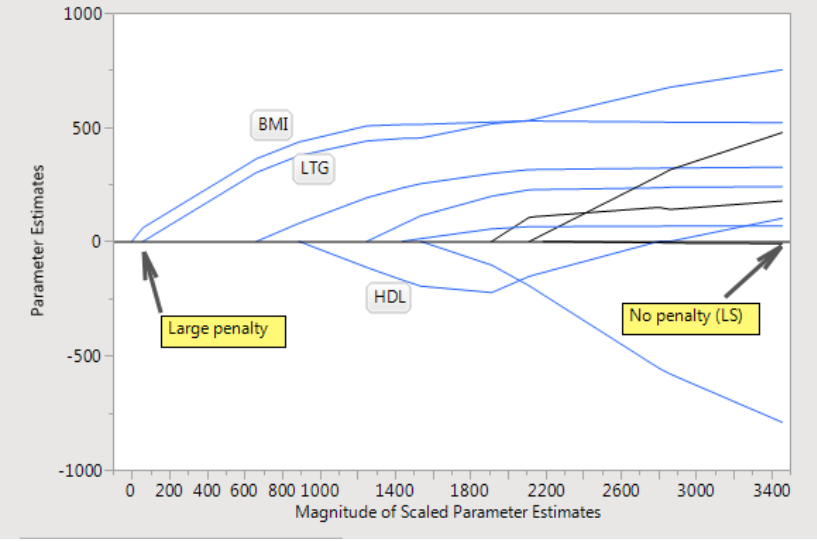
- Some DS theory was extended to the Lasso.

Demonstration

Diabetes Data Again

CV and Tuning
The Solution Path

- The Solution Path summarizes the sequence of fits.



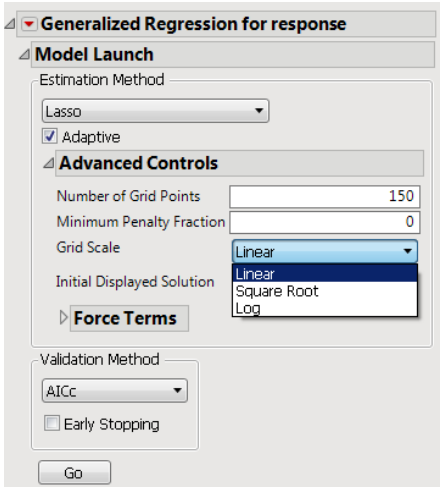
- BMI and LTG are the first two terms to enter the model.
- HDL enters with a negative coefficient, but later becomes positive as the penalty is relaxed.

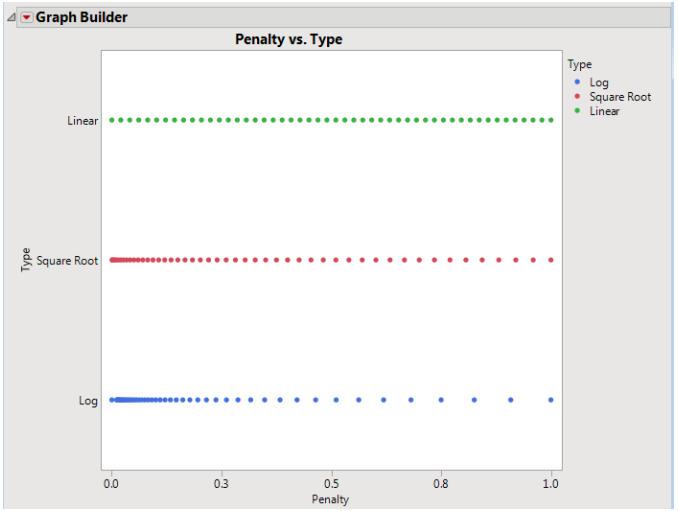
- Each line is a model parameter, the penalty decreases from left to right.

CV and Tuning

Grid of Tuning Parameters

- We have to define the grid of tuning parameters $[\lambda_1, \lambda_2, \cdots, \lambda_g]$.
  1. A linear grid gives more points near the intercept-only model.
  2. Log grid gives more points near the unpenalized model.
  3. Square root is a compromise between the two. (Default)





CV and Tuning

Cross Validation

- Cross validation refers to the process of breaking our data into *training* and *validation* sets.
  - We use the training set to estimate model parameters.
  - Take the model fit on the training set and apply it to the validation set. This gives us an idea of how it will do on new data.
  - Keep the model that fits best on the validation set.

- Most simple (most common?) case is to have a single training set and a single validation set.

CV and Tuning

Cross-Validation

Suppose we have 150 observations. We could use the first 100 for training and last 50 for validation.

Use the training piece for estimation giving us $\hat{\beta}(\lambda)$ for each λ.

For each λ, calculate SS for the validation set:

$$SSV(\lambda_j) = \sum_{i=101}^{150} \left(y_i - \boldsymbol{x}_i \hat{\beta}(\lambda_j)\right)^2$$

λ that minimizes SSV is our "best" model.

$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ \vdots & & \vdots \\ x_{100,1} & \cdots & x_{100,p} \\ x_{101,1} & \cdots & x_{101,p} \\ \vdots & \ddots & \vdots \\ x_{150,1} & \cdots & x_{150,p} \end{bmatrix} \quad \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_{100} \\ y_{101} \\ \vdots \\ y_{150} \end{bmatrix}$$

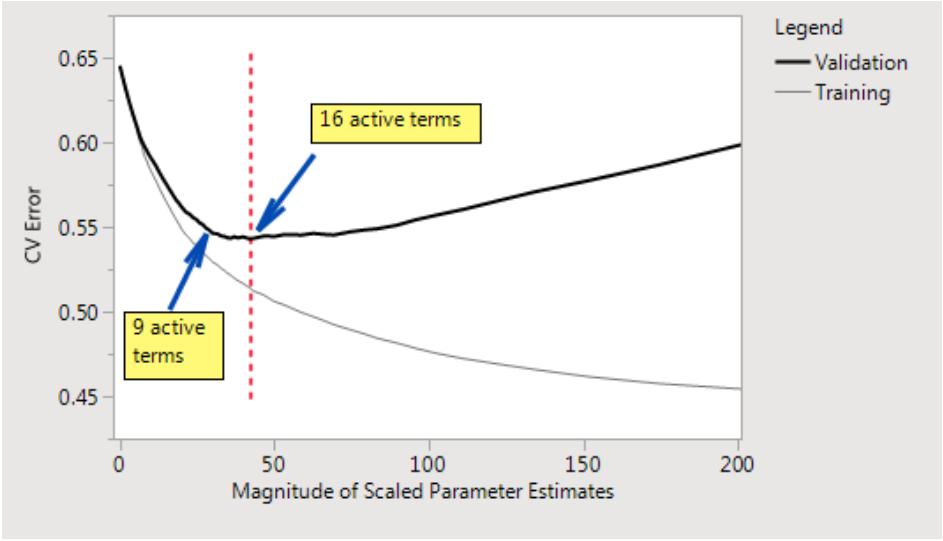
Training

Validation

CV and Tuning

Do we really want the best?

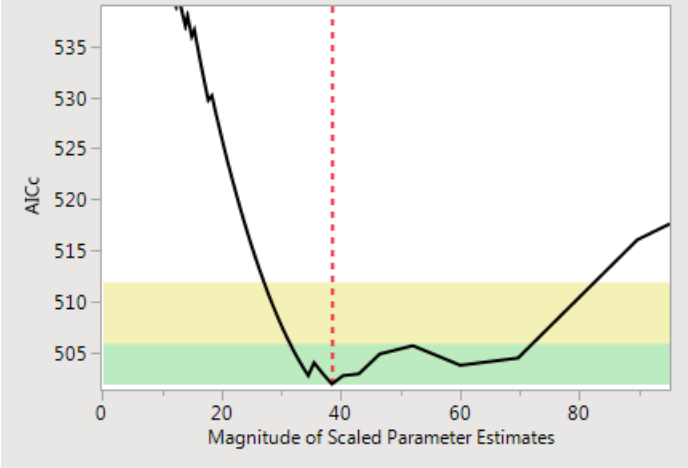
- Genreg gives you the best model in terms of AIC/CV/…
- When a simpler model is nearly as good, should we go simpler?
- If our goal is effect screening, should we go more complex?



CV and Tuning

Similar

- AIC and BIC provide guidelines regarding which models are similar.
  - AIC - Best AIC < 4 ⇒ strong evidence supporting the lesser model
  - 4 ≤ AIC – Best AIC < 10 ⇒ weak evidence for the lesser model
  - AIC – Best AIC >10 ⇒ probably avoid these models



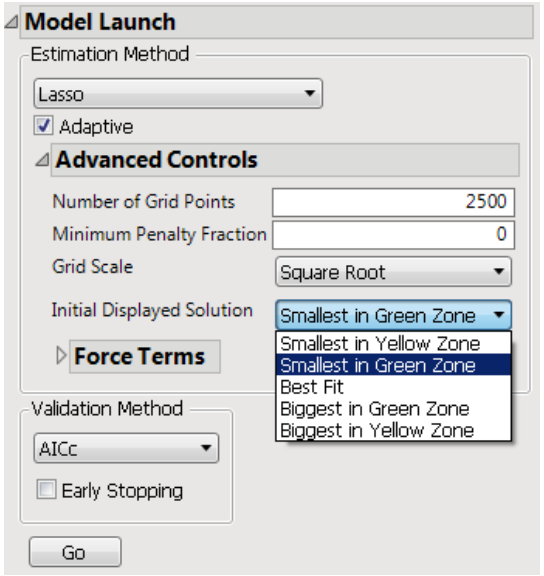
We use green and yellow zones to define these regions of similarity.

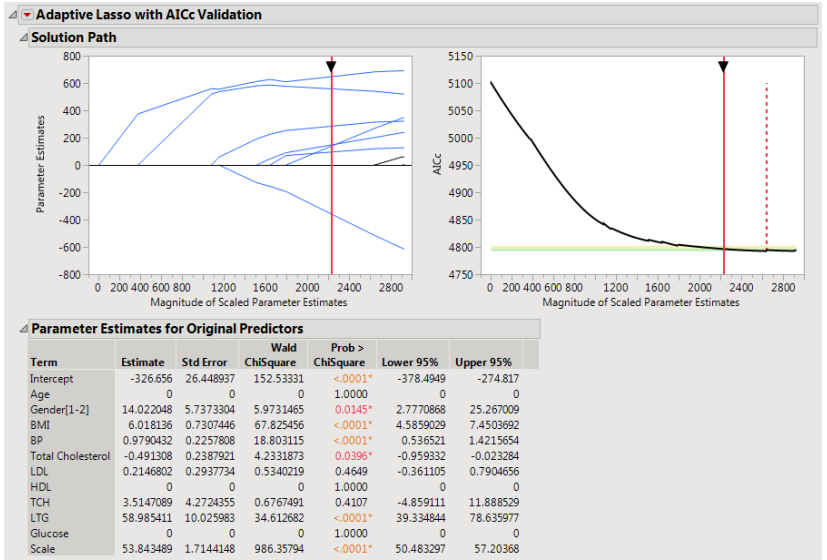
More information, see Burnham and Anderson (2003), "Model Selection and Multi-Model Inference: A Practical Information Theoretic Approach"

CV and Tuning

Similar Models

- Interactivity in Genreg makes it easy to explore models similar to the best.
- The advanced controls also provide options for starting at a model other than the best (also available in the Preferences).



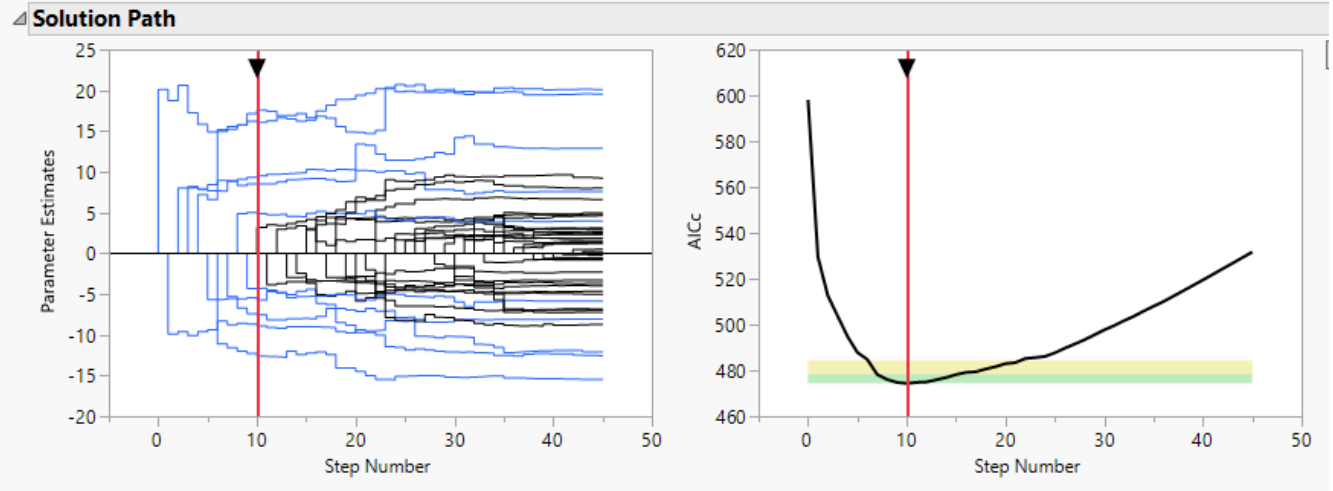


Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-326.656	26.448937	152.53331	<.0001*	-378.4949	-274.817
Age	0	0	0	1.0000	0	0
Gender[1-2]	14.022048	5.7373304	5.9731465	0.0145*	2.7770868	25.267009
BMI	6.018136	0.7307446	67.825456	<.0001*	4.5859029	7.4503692
BP	0.9790432	0.2257808	18.803115	<.0001*	0.536521	1.4215654
Total Cholesterol	-0.491308	0.2387921	4.2331873	0.0396*	-0.959332	-0.023284
LDL	0.2146802	0.2937734	0.5340219	0.4649	-0.361105	0.7904656
HDL	0	0	0	1.0000	0	0
TCH	3.5147089	4.2724355	0.6767491	0.4107	-4.859111	11.888529
LTG	58.985411	10.025983	34.612682	<.0001*	39.334844	78.635977
Glucose	0	0	0	1.0000	0	0
Scale	53.843489	1.7144148	986.35794	<.0001*	50.483297	57.20368

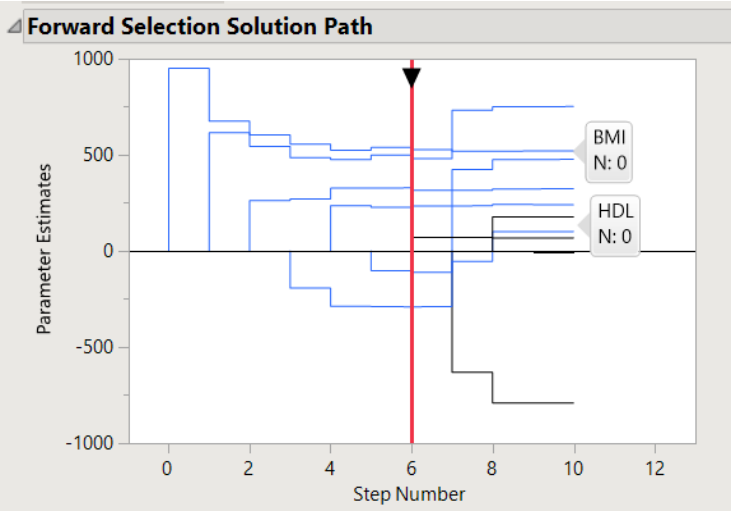
CV and Tuning

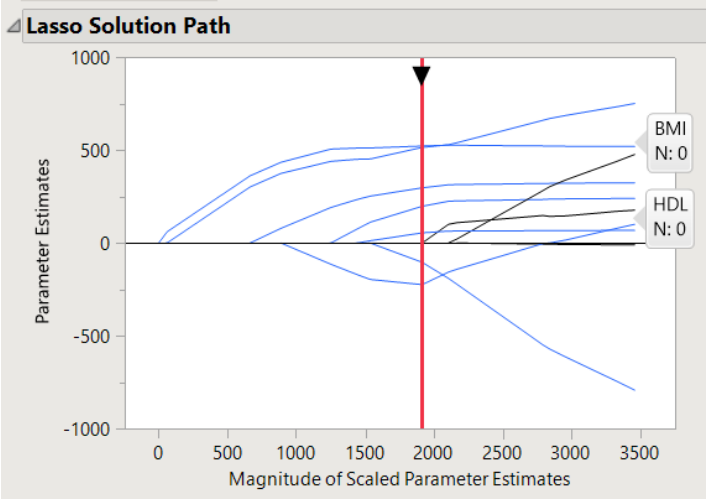
Stepwise Methods

- We've focused on penalized methods, but the idea is the exact same for stepwise methods.
- Rather than tuning a penalty value, we tune the number of steps. All the same ideas apply.



Interactive Solution Path



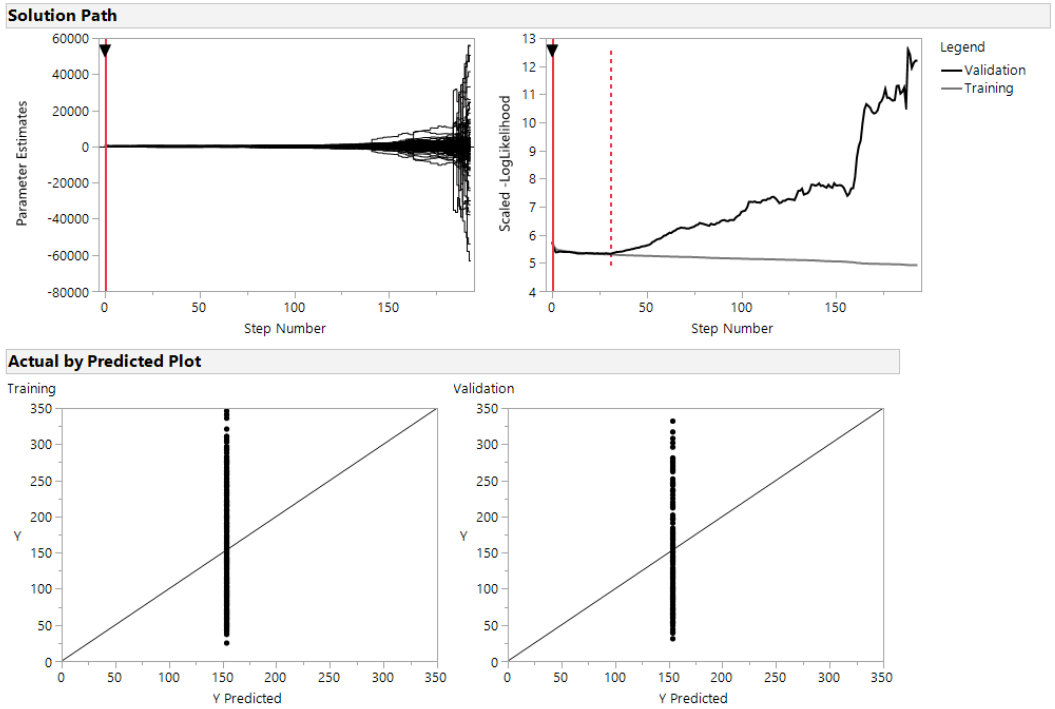


- The shapes may be different, but the information conveyed is the same regardless of method: The sequence of variables selected.

Interactive Solution Path

And Diagnostics

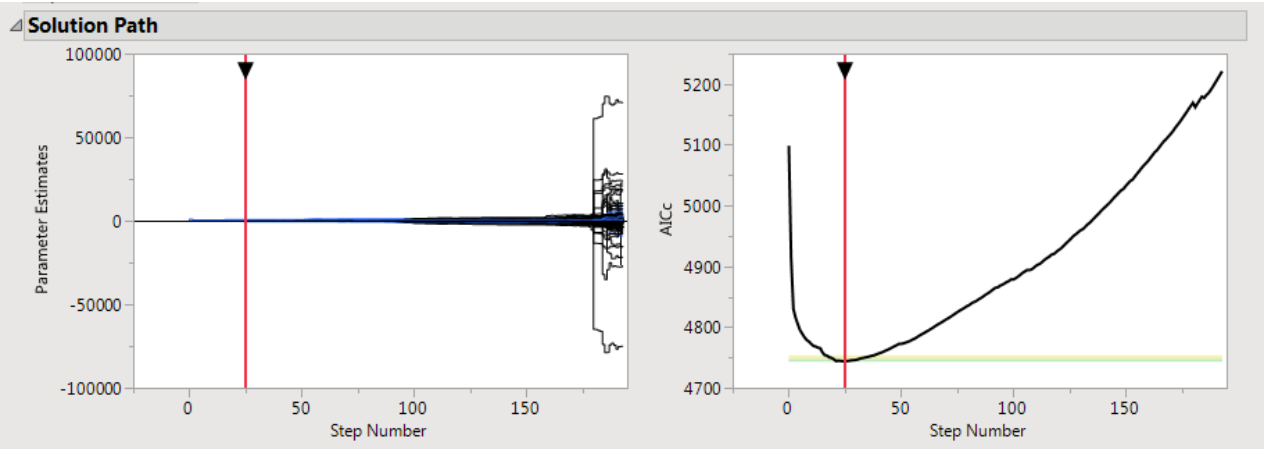
- Taking advantage of interactivity and built-in diagnostics help us understand our fits.
- Our model slows down improving on training and starts to get worse on the hold-out set.



Interactive Solution Path

Collinearity

- The solution path also gives us a clear picture of what happens when have a lot of collinearity in our predictors.
- As we add more collinearity to our model, things get unstable and our estimates blow up in magnitude towardsthe end of the path.



Interactive Solution Path

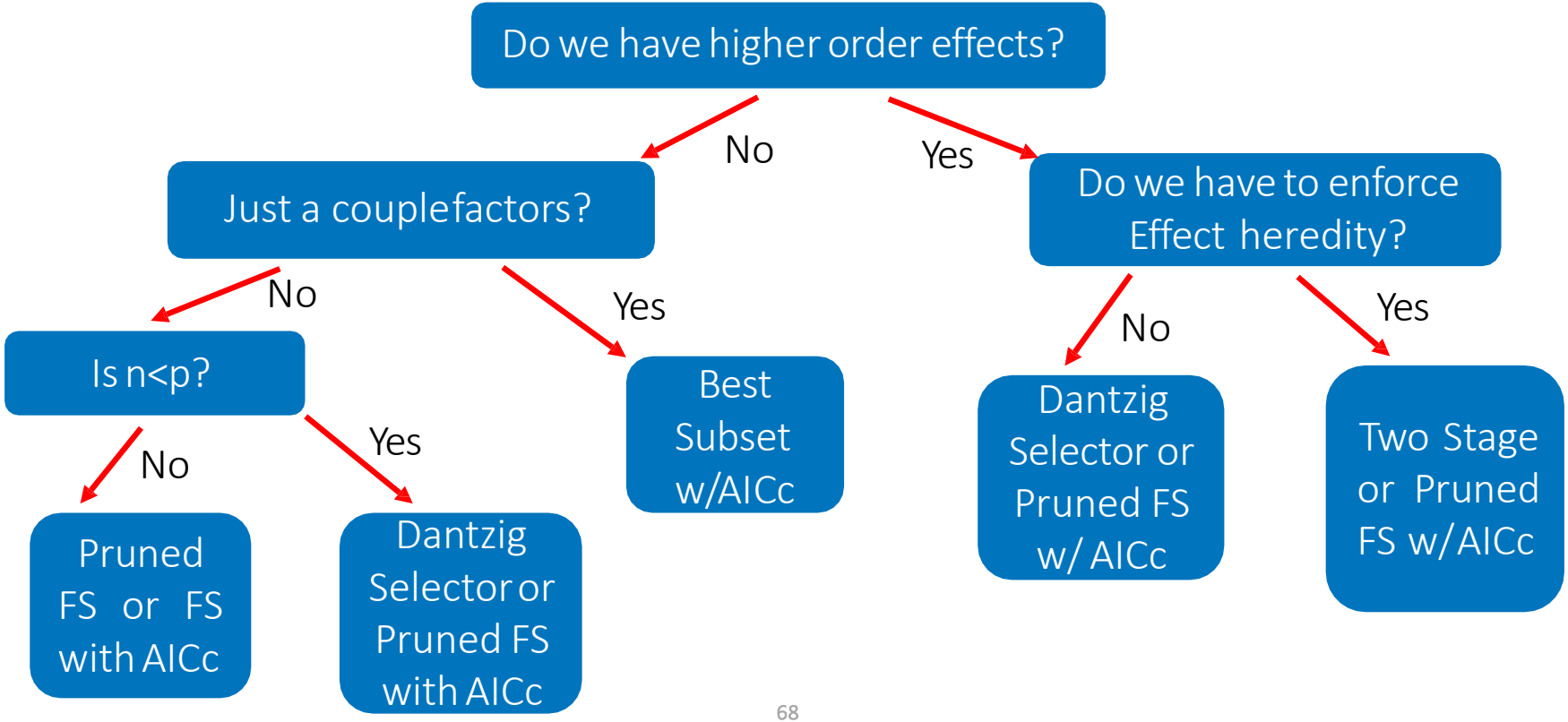
Rules of Thumb

- Here are some rules of thumb that may help…
- For designed experiments…
  - Consider Dantzig Selector, 2SFS, or Double Lasso.
  - 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 effect heredity.
- Is the normal distribution reasonable? Don't be afraid to try Gamma…

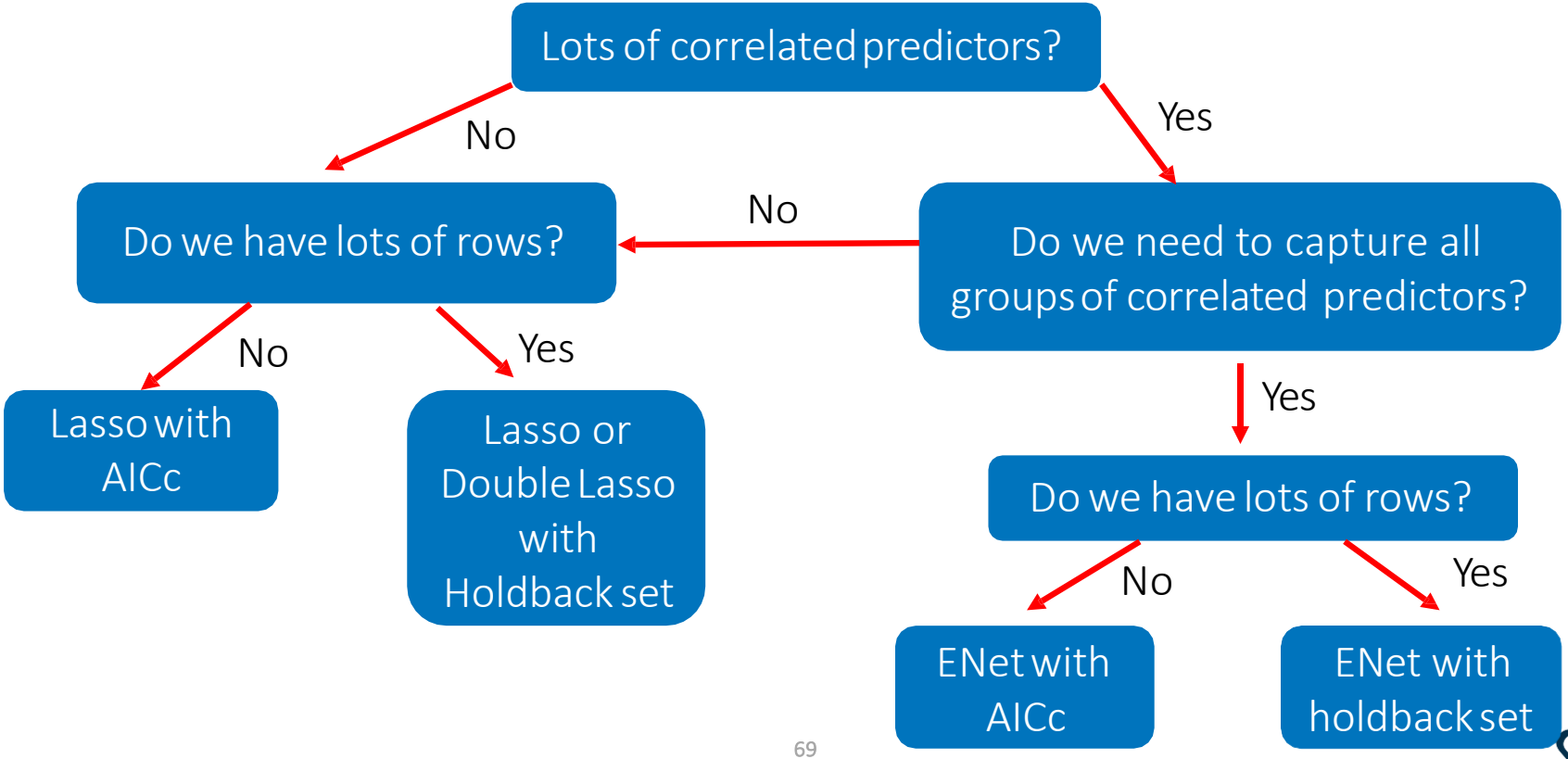
More Rules of Thumb

- We can summarize what methods to consider in a flow chart.
- BIG DISCLAIMER – These *are not* hard and fast rules for what one should do. They are merely the methods that I tend to find myself using in different circumstances.
- Can't decide between two (or more) methods? Consider trying them both and taking the intersection or union of the results.
- Break it down into two settings: experimental and observational data.

Experimental Data



Observational Data



Demonstration