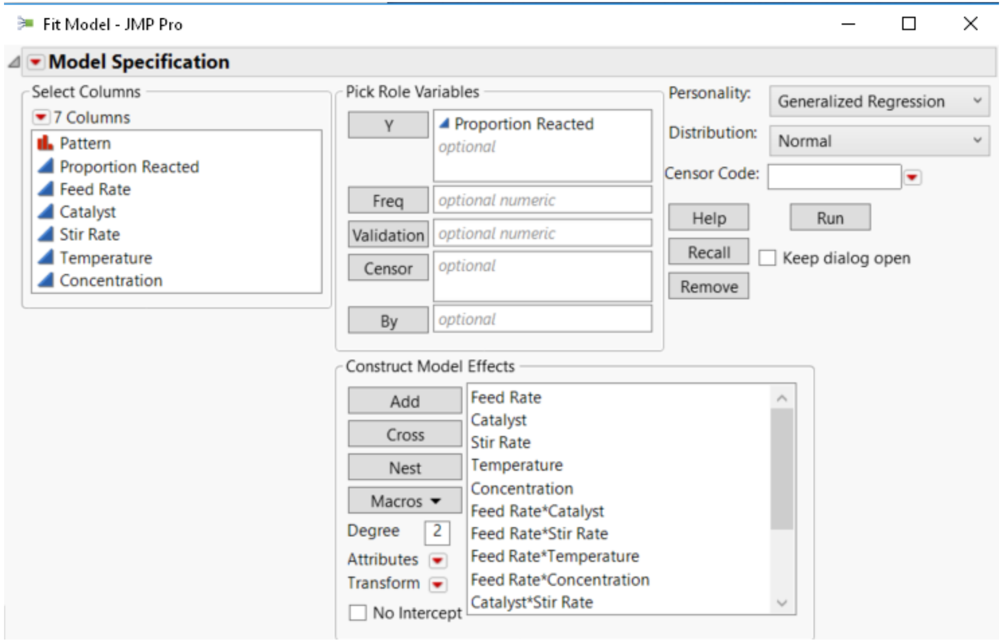
Using Generalized Regression to Analyze Observational Data

11/11/2021 Developer Tutorial

Clay Barker

Where did we leave off?

Last week, we talked about using the Generalized Regression platform in JMP Pro to analyze Designed Experiments.



What is Variable Selection Again?

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. Put them all into a variable selection procedure and see what happens.

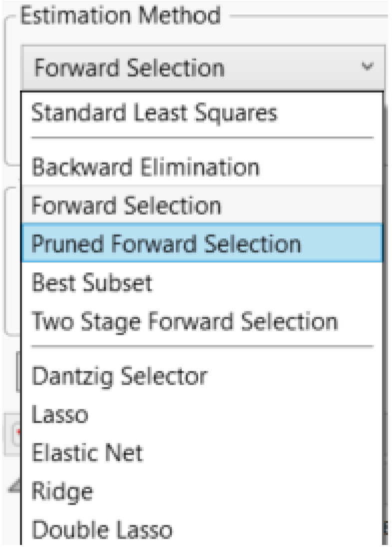
- But we still need to think carefully about our data!

Where did we leave off?

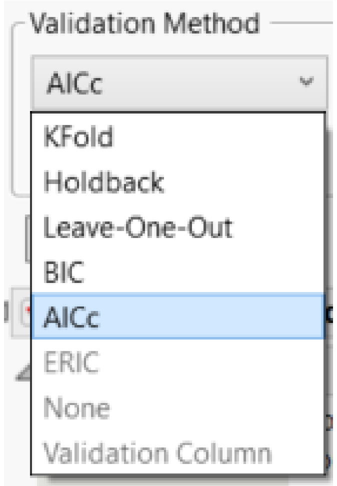
Just to recap

With experiments, we tend to stick with stepwise methods with the AICc.

And we tend to request Effect Heredity.



Stepwise Methods



AICc and BIC

The same approach may not be optimal for observational data.

1. We probably have lots more data.
2. We almost certainly don’t have orthogonality.
3. We may be more interested in prediction than interpretation.

Today we’ll talk about the methods in Genreg that we tend to recommend for observational data sets.

Estimation

How good is an estimator?

Whenever we estimate something, how do we measure how good it is?

There are two things to consider?

1. How far from the truth do the estimates tend to be? (Bias)
2. How variable are our estimates? (Variance)

We combine the two to define the Mean Square Error of an estimator.

Mean Squared Error$(\hat{\theta})$ = Bias$\left(\hat{\theta}\right)^2$ + Variance$(\hat{\theta})$

Estimation

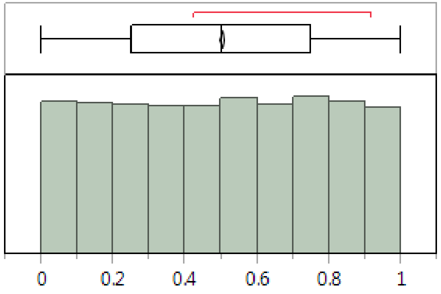
An exaggerated example

Suppose we do a simulation to compare two estimators.

- Estimator 1 is centered at the truth (.5), but highly variable.
- Estimator 2 is slightly biased, but much less variable.

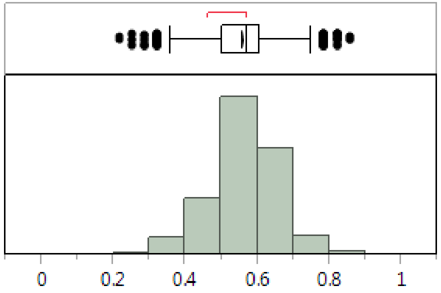
We'd almost certainly prefer Estimator 2, right?

Estimator 1



Summary Statistics	
Mean	0.5003234
Std Dev	0.2883447
Std Err Mean	0.0028834
Upper 95% Mea	0.5059755
Lower 95% Mean	0.4946713
N	10000

Estimator 2



Summary Statistics	
Mean	0.5548714
Std Dev	0.0876164
Std Err Mean	0.0008762
Upper 95% Mea	0.5565889
Lower 95% Mean	0.553154
N	10000

Estimation

Ordinary Least Squares

Often when we think of regression, we think of least squares estimation

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

The Gauss-Markov theorem tells us that $\hat{\beta}_{OLS}$ has the minimum variance of all unbiased estimators.

...but OLS estimates can have high variance.

...in particular when our predictors are highly correlated.

Penalized Regression

Maybe some bias is OK?

High variance in OLS estimates can make our model not fit new data well.

Maybe a biased but less variable estimator would generalize better?

This is the motivation behind *ridge regression*.

Penalized Regression

Ridge Regression

Hoerl and Kennard (1970) proposed ridge regression.

Instead of OLS, what if we minimize a penalized sum of squared errors?

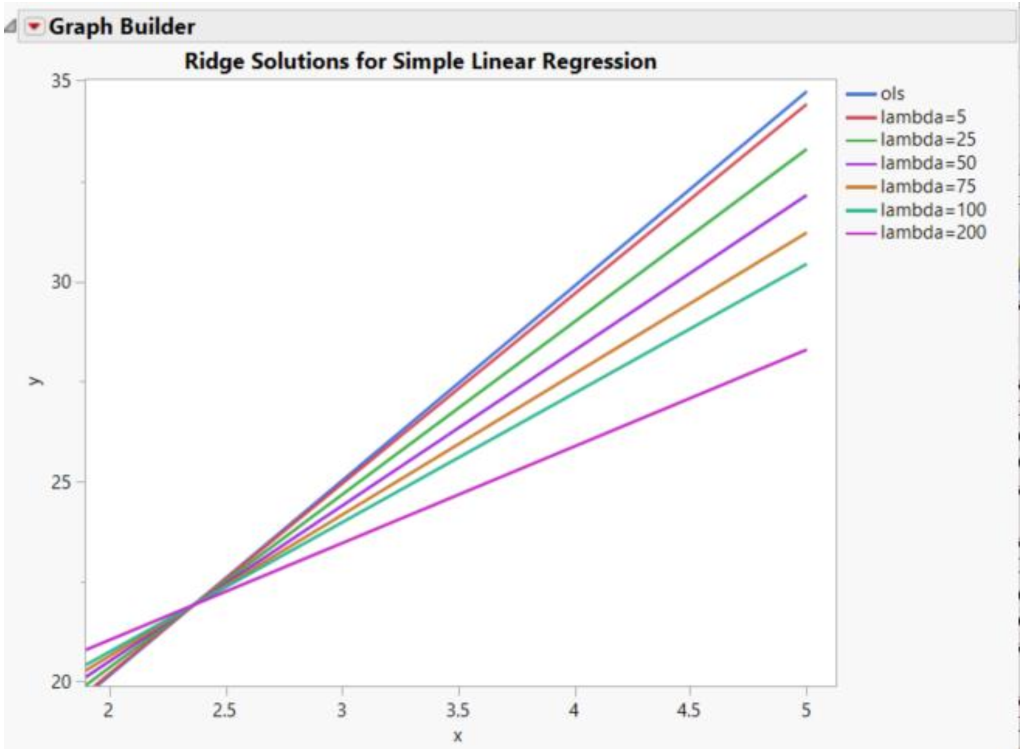
$$\hat{\beta}_{ridge} = \arg\min_{\beta} \sum_i (y_i - x_i\beta)^2 + \frac{\lambda}{2}\sum_j \beta_j^2$$
$$= \left(X^T X + \lambda I_p\right)^{-1} X^T y$$

λ is a *tuning parameter* that controls the magnitude of parameters.

- $\lambda = 0$ is the usual OLS solution.
- As λ increases, parameter estimates move toward zero. Shrinkage!
- Stabilizes estimates when predictors are highly correlated.

Penalized Regression

Ridge Regression

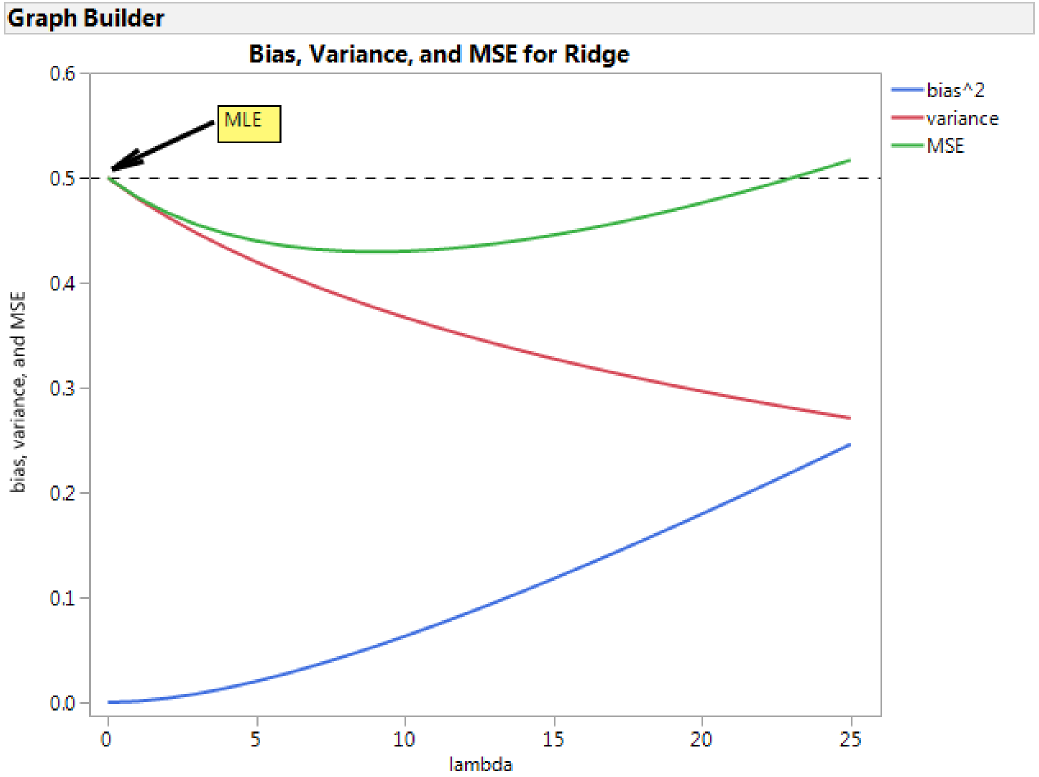


Easiest to see how this works with a single predictor.

As the tuning parameter increases, the slope of the fitted line shrinks to zero.

Penalized Regression

Can Ridge have a lower MSE than OLS?



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

Penalized Regression

Choosing the Tuning Parameter

In order to beat OLS, we need to carefully choose the tuning parameter λ.

How should we do that?

Define a grid of values $[\lambda_1, \lambda_2, \cdots, \lambda_k]$

Try out each value of λ and see which one fits the best (AICc, BIC, CV).

Usually we'd choose $\lambda_1 = 0$ to include OLS.

Very similar to what we talked about last week with stepwise methods:

Fit a sequence of models and keep the best.

Penalized Regression

The Importance of Ridge Regression

At the 2021 Joint Statistical Meetings, Trevor Hastie of Stanford had a talk celebrating the 50th anniversary of ridge regression.

Ridge or more formally L2 regularization shows up in many areas of statistics and machine learning. It is one of those essential devices that any good data scientist needs to master for their craft.

Ridge Regularization: An Essential Concept in Data Science

Trevor Hastie
Department of Statistics
Department of Biomedical Data Science
Stanford University

Penalized Regression

A Family of Models

Ridge opened the door to a variety of penalized regression 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

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

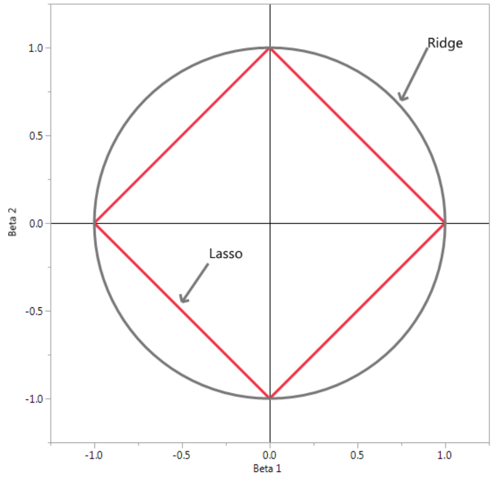
The absolute value penalty is a pain compared to ridge.

Penalized Regression

Ridge and Lasso Geometry

Instead of thinking about penalizing the SSE, we can think about these methods as constrained optimization problems.

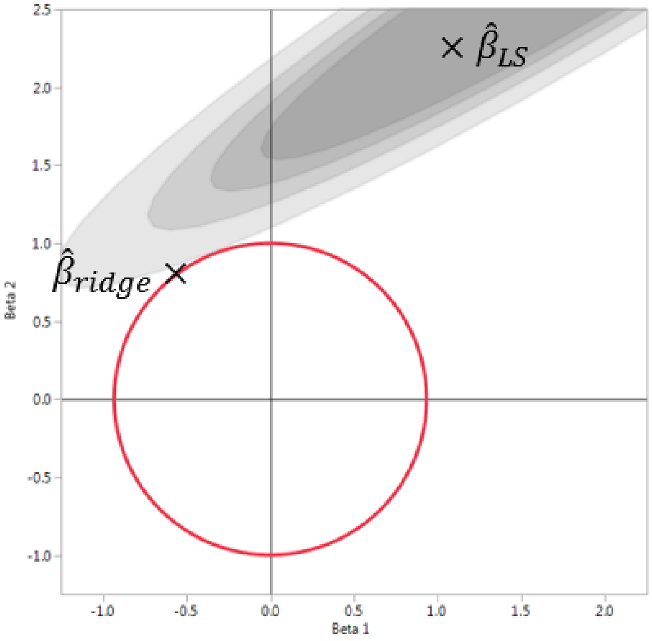
- Lasso: $\min_{\beta} \sum_i (y_i - x_i\beta)^2$ such that $\sum_j |\beta_j| \leq s$
- Ridge: $\min_{\beta} \sum_i (y_i - x_i\beta)^2$ such that $\sum_j \beta_j^2 \leq s$

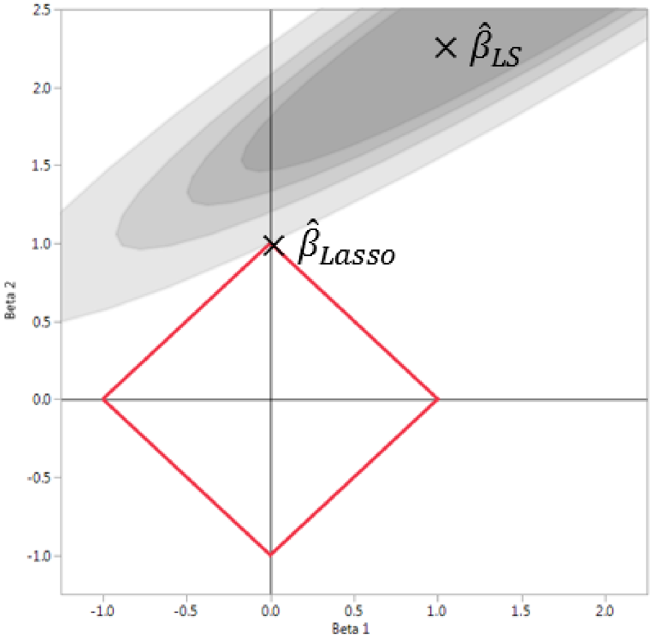


In two dimensions, the feasible regions for lasso and ridge are a diamond and circle respectively.

Penalized Regression

More Geometry





Corners on the lasso feasible region allow for intersections at zero (selection).

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

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 vs Lasso

Example

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

Elastic Net "stretches" to select groups of correlated variables.

Which solution is better? It depends.

Lasso will be simpler and probably predict well.

Elastic Net may have a more meaningful interpretation.

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

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.

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

Penalized Regression

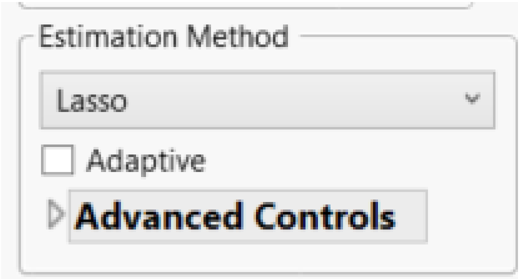
Adaptive Lasso

But be careful! If OLS is unstable, the adaptive lasso may stink.

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



My advice: proceed with caution.

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.

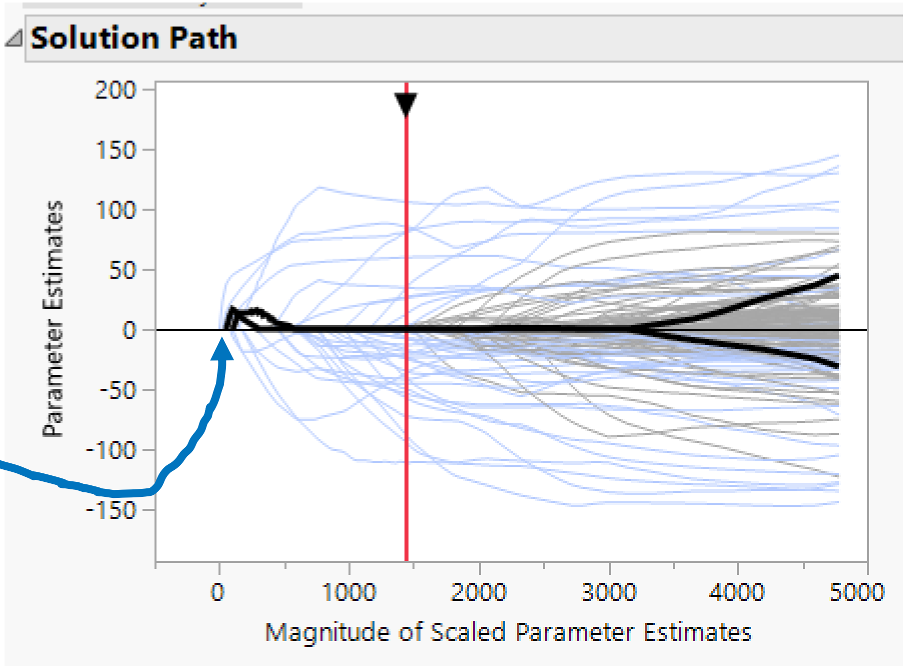
Penalized Regression

Double Lasso

When will the second pass of the lasso pay off the most?

…if variables come and go before the best solution in the first pass.

If your Lasso solution path looks like this, then the double lasso may fit slightly better.



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.

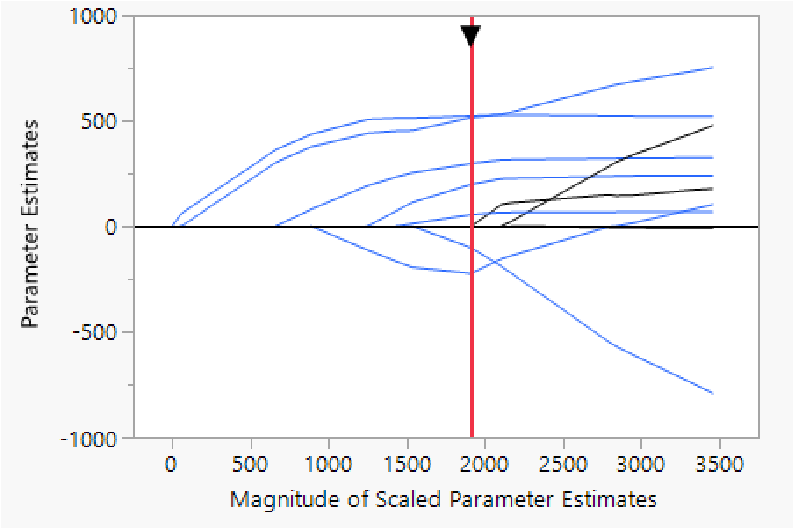
$$\hat{\beta}_{DS} = \arg\min_{\beta} \sum_j |\beta_j| \text{ subject to } |X^T(y - X\beta)|_\infty \leq s$$

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

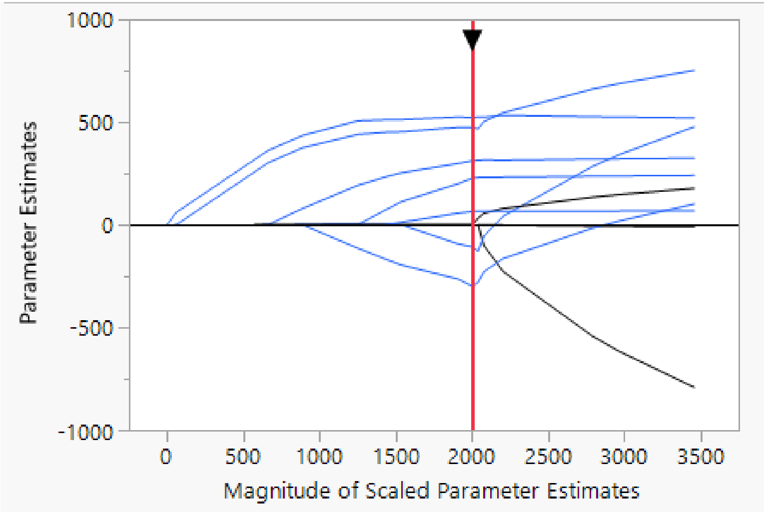
This is a penalized regression technique, but it is mainly recommended for analyzing designed experiments.

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

The Dantzig Selector

From Efron, Hastie, and Tibshirani (2007)

From our brief study, the inherent criterion in DS for including predictors in the model appears to be counterintuitive, and its prediction accuracy seems to be similar to that of the Lasso in some settings, and inferior in other settings. Hence we find little reason to recommend the Dantzig selector over the Lasso.

Might be worth trying with modeling the results of a designed experiment, but skip it for observational data.

Some Options to Consider

Some Options to Consider

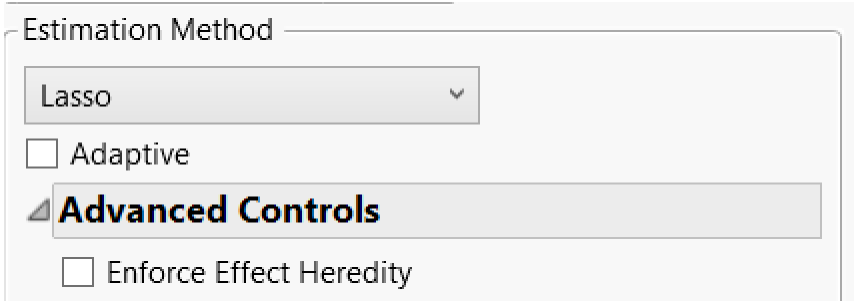
Effect Heredity

Recall: *Effect Heredity* means that in order for A*B to be in the model, A and B must also be in the model.

Stepwise Methods accommodate heredity very easily.

Penalized Methods? Not so much.

We will try if you request it. But if heredity is truly important, best to stick with stepwise methods.



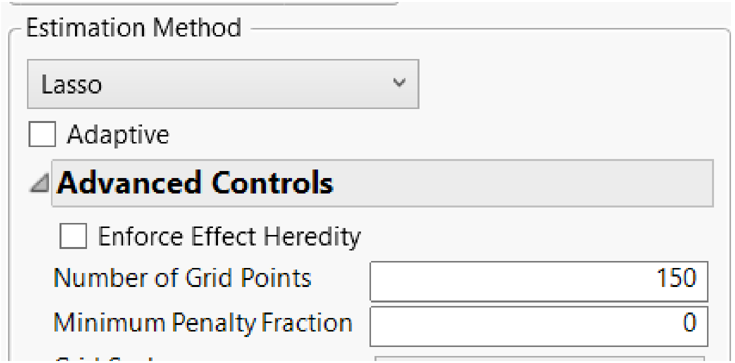
Some Options to Consider

Grid Controls

Recall that when we fit a penalized regression model, we evaluate over a grid of tuning parameters $[\lambda_1, \lambda_2, \cdots, \lambda_k]$.

$k = 150$ by default

We calculate λ_k. It is the smallest value that zeroes out all of the coefficients.



And $\lambda_1 = a\lambda_k$, you can specify $a \in [0,1)$.

In the Genreg controls, we call this the “Minimum Penalty Fraction”.

Some Options to Consider

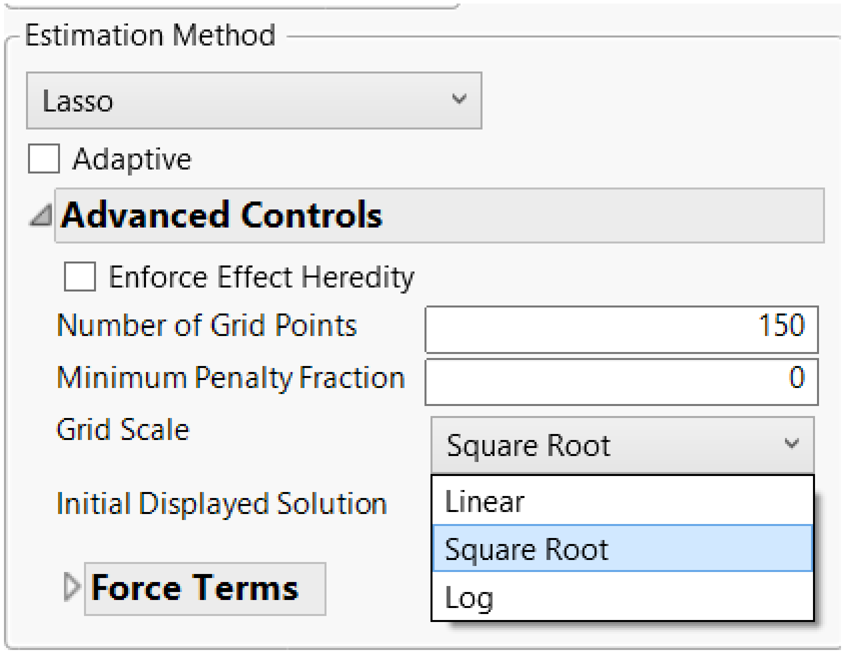
Grid Scale

We just saw how we choose the minimum and maximum grid points. The *Grid Scale* lets us choose how to choose the points in between.

Linear: lots of models early in path

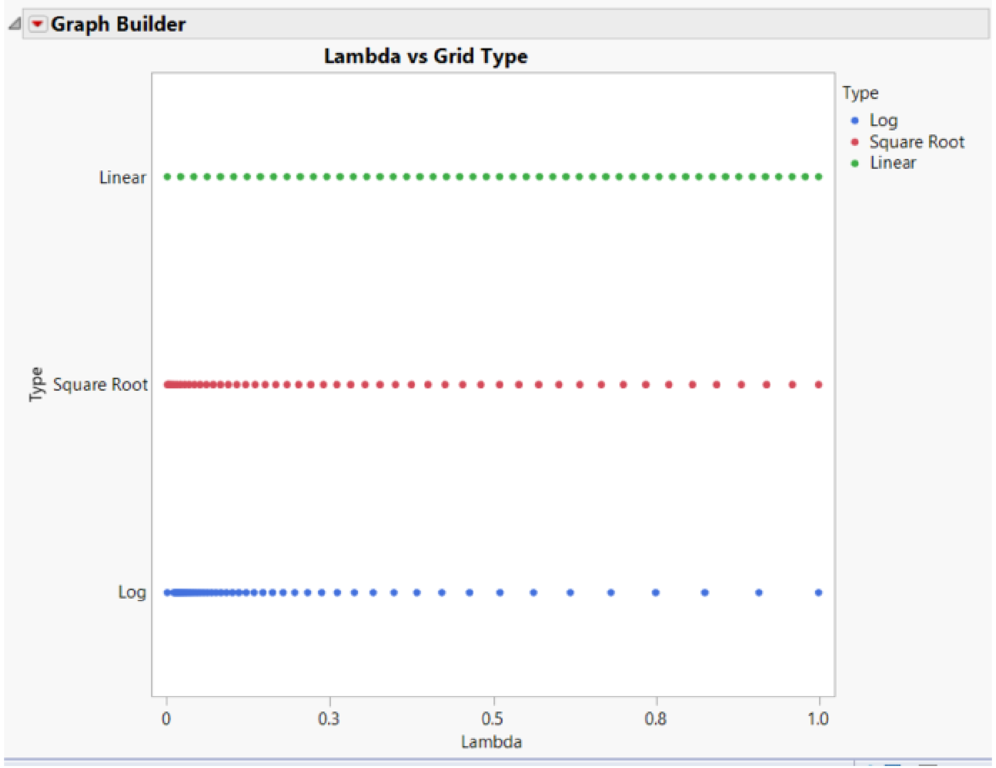
Log: lots of models late in the path

Square Root: A great in-between



Some Options to Consider

Grid Types



Grid points closer to 0 are closer to the unpenalized fit.

Grid points closer to 1 are closer to the intercept only model.

Some Options to Consider

Forced Terms

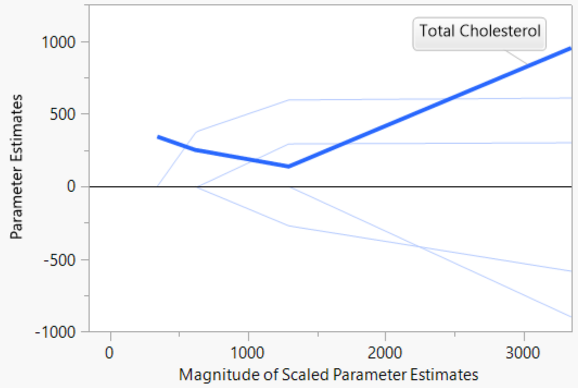
As advertised, Forced Terms are omitted from the penalty. So they are in every model in the solution path.

Force Terms

Forced terms are not included in the penalty.

Force	Term
✓	Intercept
	BMI
	BP
✓	Total Cholesterol
	LDL
	HDL

Solution Path



Parameter Estimates for Original Predictors

Term		Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare
Intercept		-97.73917	25.923427	14.215185	0.0002*
BMI		6.5633035	0.7476607	77.061221	<.0001*
BP		1.0367349	0.2271829	20.824962	<.0001*
Total Cholesterol	Forced in	1.3633829	0.2426358	31.573733	<.0001*
LDL		-1.470191	0.2779251	27.982843	<.0001*
HDL		-2.199638	0.2773035	62.920446	<.0001*

Some Options to Consider

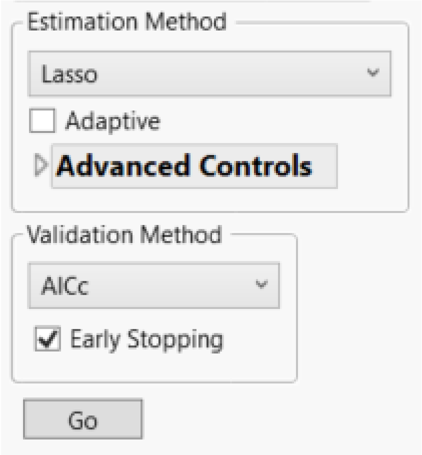
Early Stopping

For very large problems, it might make sense to try *Early Stopping*.

What exactly does that mean?

If we go 10 steps after the best fit,

we stop instead of going through the entire grid.



<u>Example</u>

Lets say λ_j provides the best fit so far.

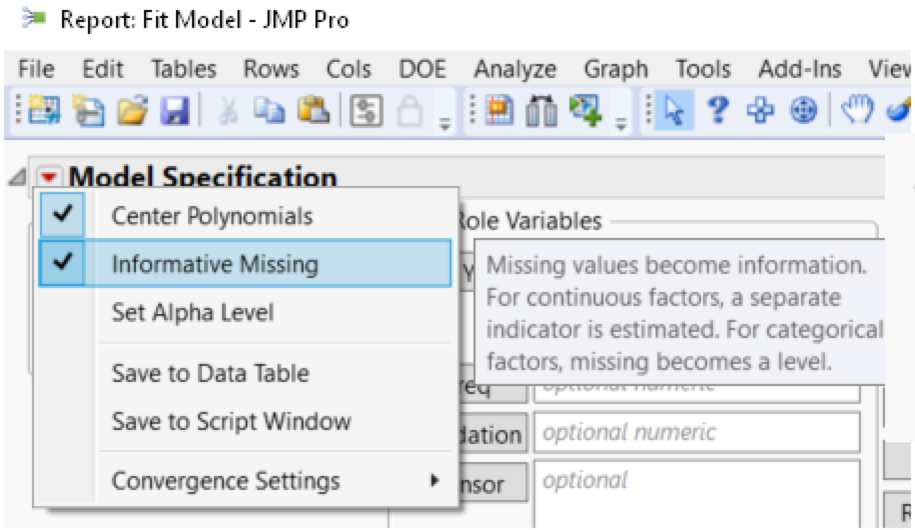
If we get to λ_{j+10} and λ_j is still the best, we go ahead and stop.

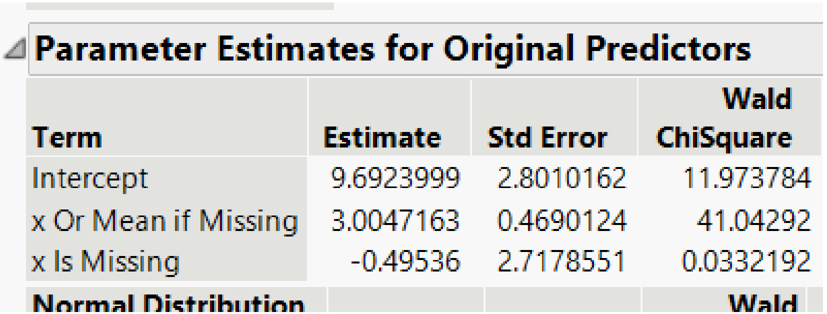
Of course sometimes we end up stopping **too soon**, so use caution.

Some Options to Consider

Informative Missing

If we have missing values in any of our predictors, we may want to consider the *Informative Missing* option in the Fit Model launch.





Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare
Intercept	9.6923999	2.8010162	11.973784
x Or Mean if Missing	3.0047163	0.4690124	41.04292
x Is Missing	-0.49536	2.7178551	0.0332192

Some Options to Consider

Informative Missing

	x	y
1	3	17.9
2	7	29.9
3	8	33.3
4	10	39.9
5	•	19.9
6	2	16.1
7	6	28.6
8	•	20.0
9	3	18.3
10	1	13.0
11	7	31.5
12	•	34.8

Original Data

x Or Mean If Missing	x Is Missing	y
3	0	17.9
7	0	29.9
8	0	33.3
10	0	39.9
5.22	1	19.9
2	0	16.1
6	0	28.6
5.22	1	20.0
3	0	18.3
1	0	13.0
7	0	31.5
5.22	1	34.8

Modified Data

When you ask for informative missing, we essentially convert it to the modified data for modelling.

That way we don't have to drop any rows.

Some Options to Consider

Initial Solution

By default, we give you the best fitting model.

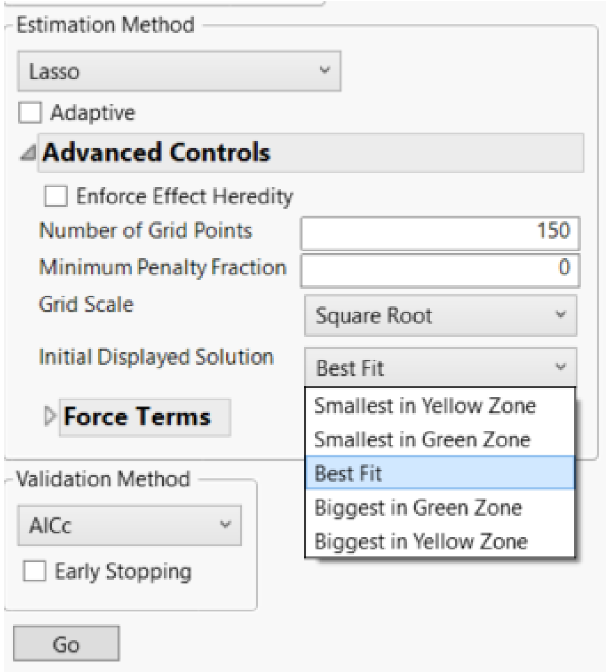
But we can give you slightly bigger or smaller models that are still supported by the data.

Let γ be the best AICc or BIC.

Green Zone: [γ, γ +4]

Yellow Zone: [γ +4, γ +10]

Works similarly with k-fold.



Some Options to Consider

…or maybe not

Do we really need to be concerned with these Advanced Controls???

We’ve chosen defaults carefully, so probably not often.

But they’re there if you need extra care with non-standard problems.

Summary

The Generalized Regression platform is the place to build regression models.

...for both designed experiments and observational data.

Some things to keep in mind...

- Penalized regression shows great promise for observational data.
- Effect Heredity probably isn't necessary.
- Use a hold-out set if you have enough data, otherwise the AICc.

Thanks!

Clay.Barker@sas.com

sas.com