PENALIZING YOUR MODELS

AN OVERVIEW OF THE GENERALIZED REGRESSION PLATFORM

Michael Crotty & Clay Barker
Research Statisticians
JMP Division, SAS Institute

OUTLINE

- Motivation
- Generalized Linear Models
- Penalization Methods
- New features
  - 11.1
  - 11.2
  - 12
- Examples
  - Heart disease data
  - Bowl game data
- Conclusions
- References

- Basic regression model:

$$E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots$$

- Why do we build regression models?
  1. To better understand our data and make decisions
  2. To make predictions for new observations

- The Generalized Regression platform in JMP Pro can help users with both of these goals.

MOTIVATION

- The Generalized Regression platform (personality of Fit Model) allows users to fit Penalized Generalized Linear Models.
  - Penalized means that we are using a penalized likelihood to fit our model. This allows us to do variable selection and shrinkage.
  - Generalized means that our response doesn't have to be normally distributed. Not everything is normal: counts, yes/no, skewed, outliers, …

GENERALIZED LINEAR MODELS

- Many regression tools assume the response variable is approximately normally distributed.
- But, this isn't always an appropriate assumption…
  - Insurance claims are often skewed (gamma distribution)
  - Presence of heart disease is yes/no (binomial distribution)
  - Number of horse kick victims (Poisson distribution)
- JMP 11 supports 7 distributions for the response.
- JMP 12 supports 14 distributions for the response.

RESPONSE DISTRIBUTIONS

The following distributions are available in the Generalized Regression platform.

Continuous	Discrete	Zero-inflated
Normal	Binomial	***ZI Binomial***
Cauchy	***Beta Binomial***	***ZI Beta Binomial***
Exponential	Poisson	ZI Poisson
Gamma	Negative Binomial	ZI Negative Binomial
Beta		***ZI Gamma***

Note: Distributions added in JMP 12 are ***italicized*** above.

BRIEF DETAILS

- Linear model: $E(\mathrm{Y}) = \mu$ where $\mu = \mathrm{X}\beta$
- Generalized linear model (in three parts):
  - Mean component: $E(\mathrm{Y}) = \mu$
  - Systematic component: linear predictor $\eta = \sum_1^p \mathrm{x}_j \beta_j$
  - Link function: $\eta = g(\mu)$
- Then, $y|x, \beta$ is *f* distributed with link function *g*.
- Most common GLM is probably logistic regression.

BENEFITS

- A penalty is not a bad thing; rather, it's a great thing!
- Maximum likelihood estimation gives us the model that best fits the *observed* data.
- If we optimize a penalized likelihood instead,
  - We will predict better on new data
  - We should have a model that is easier to interpret
  - We can overcome problems with our data:
    - Not enough observations (more columns than rows, "$n<p$")
    - Correlated predictors (multicollinearity)

THE PENALTY

- We want our model to fit well, but not get too complex.
- Objective function: $Q(\beta) = -log\, L(\beta) + \lambda \sum_{j=1}^{r} p(\beta_j)$
- λ controls the stiffness of the penalty.

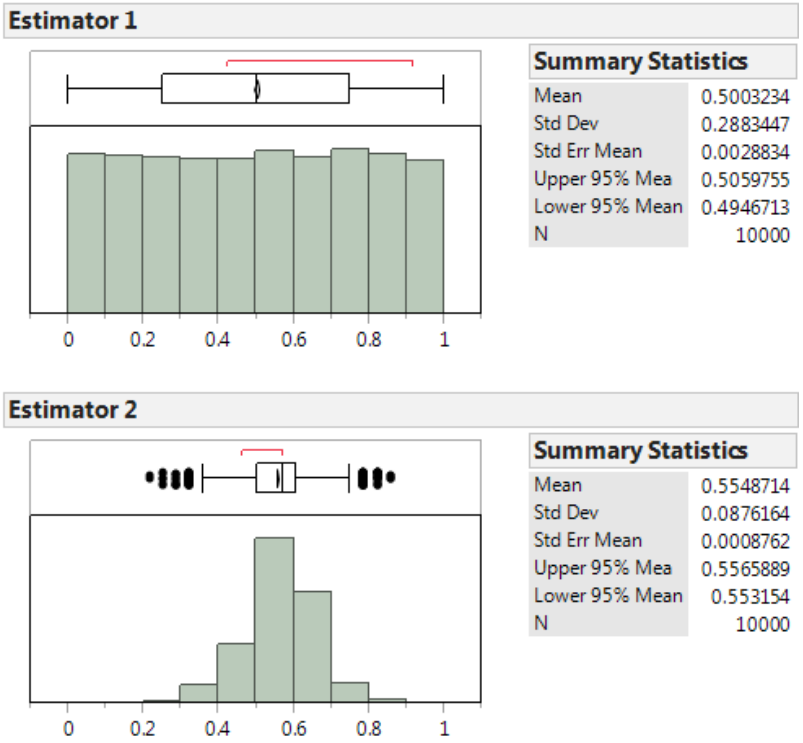
Method	$p(\beta_j)$
Lasso	$\lvert\beta_j\rvert$
Ridge	β_j^2
Elastic Net	mix of $\lvert\beta_j\rvert$ and β_j^2

- Many penalties have been proposed, but these three are available in JMP Pro.

PENALIZATION METHODS

KEY IDEA

- What is special about these penalized estimates?
- Key idea: accept some bias in order to reduce variance.
- Figure: Two different estimators of Pr(coin lands on heads)



OVERFITTING

- What if we have tons of predictors?
  - Don’t use them all!
- Overfitting means that our model will fit our observed data very well, but it will perform poorly on new observations.
  - The Lasso and Elastic Net shrink some coefficients to zero, providing us with variable selection.
  - Selection and shrinkage help us avoid overfitting.

HOW TO CHOOSE AN ESTIMATION METHOD?

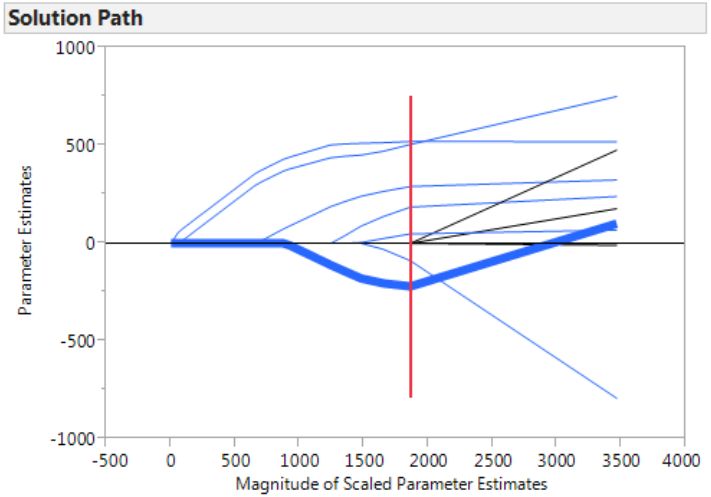
		Selection	
		No	Yes
Shrinkage	No	ML	Forward Selection
	Yes	Ridge	Lasso & Elastic Net

- Lasso tends to give you a more parsimonious model than Elastic Net.
- Elastic Net can better handle collinearity.

PENALIZATION METHODS

SOLUTION PATH GRAPH

- The solution path is crucial to understanding what has happened in the model fitting process.
  - Each line in the plot represents the value of a parameter in the model.
  - As the penalty is relaxed more terms enter the model (left to right).
  - MLE estimate is at the far right of the solution path (in most cases).
  - Parameter estimates can change sign as they travel along the path.

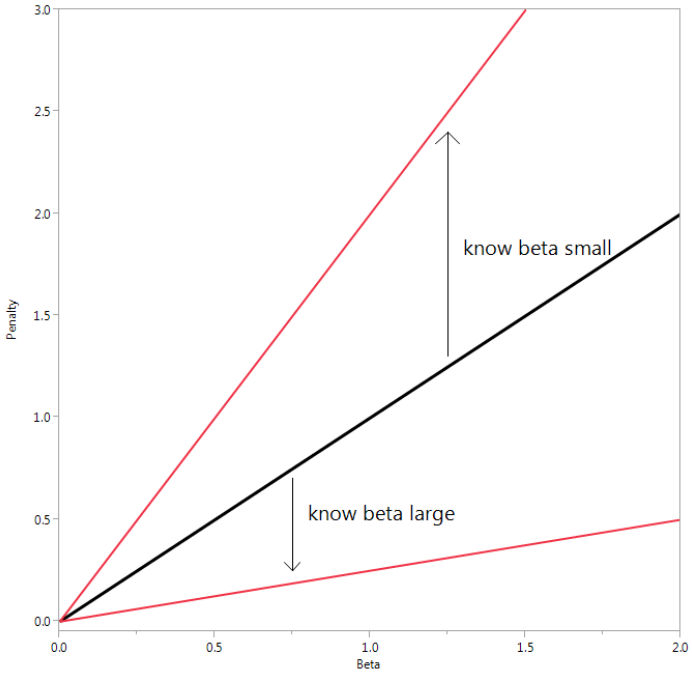


Parameter Estimates for Centered and Scaled Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	152.13348	2.5576133	3538.1795	<.0001*	147.12065	157.14631
Age	0	0	0	1.0000	0	0
Gender[1]	185.24066	58.529686	10.016598	0.0016*	70.524583	299.95674
BMI	520.77896	68.302972	58.133659	<.0001*	386.90759	654.65032
BP	290.64335	65.574768	19.644784	<.0001*	162.11916	419.16753
Total Cholesterol	-88.78233	70.988674	1.5641374	0.2111	-227.9176	50.352918
LDL	0	0	0	1.0000	0	0
HDL	-219.8571	65.854845	11.14565	0.0008*	-348.9302	-90.78395
TCH	0	0	0	1.0000	0	0
LTG	505.65164	81.675706	38.328078	<.0001*	345.57019	665.73308
Glucose	48.541624	63.311866	0.587839	0.4433	-75.54735	172.6306
Scale	53.77074	1.6381454	1077.4252	<.0001*	50.560034	56.981446

ADAPTIVE LASSO AND ELASTIC NET

- What if we know which predictors are good and/or bad?
- The ℓ_1 penalty can be adaptive.
  - The adaptive forms of the Lasso and Elastic Net penalties use weights generated from the original Maximum Likelihood estimates.



VALIDATION METHODS

- The platform has many options for validation methods. These are used for finding the optimal value of the penalty.
  - Kfold
  - Holdback
  - Leave-One-Out
  - BIC (not available for Ridge)
  - AIC (not available for Ridge)
  - None (only available for Maximum Likelihood)
  - Validation Column
- Note: Maximum Likelihood can only use the last two methods (no penalty to optimize).

NEW IN JMP 11.1

- Improved computation time (sometimes drastically).
- Lines in the solution path became selectable.
- Early stopping rules were added (another way of speeding up fitting).
- Parameter Estimates for Original Predictors added to report output.
- Bug fixes.

NEW IN JMP 11.2

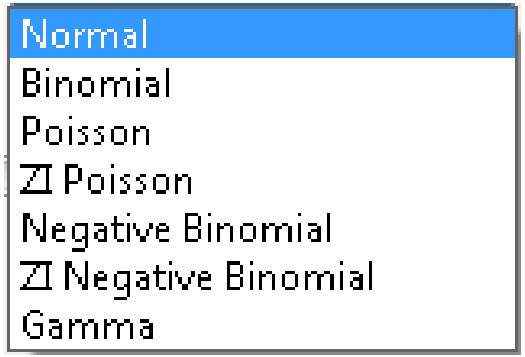
- Improved computation time.
- Better drawing of the solution path (no longer selectable on y=0 line).
- Added support for validation columns with a Test set (training/validation/test).
- Bug fixes.

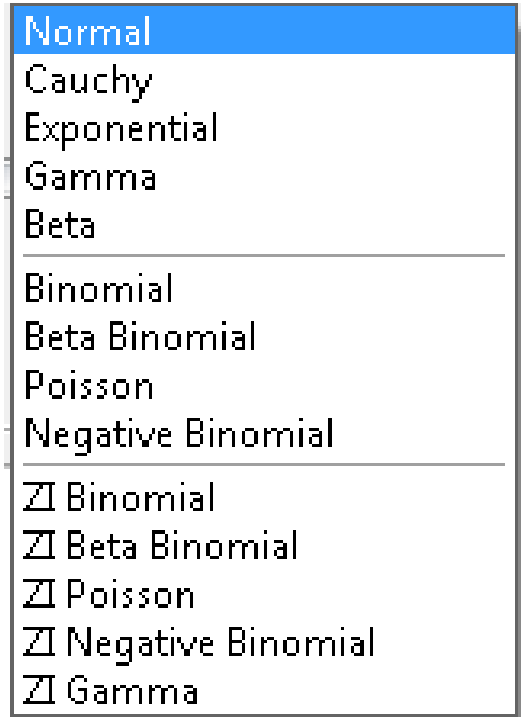
NEW IN JMP 12

- Speed (by using a new, faster optimizer for fitting models).
- New response distributions.
- Interactive solution path.
- Advanced controls for estimation options.
- New diagnostic plots.
- Inverse prediction.
- Multiple comparisons.
- Ability to force specific terms into the model.
- Forward selection.
- Bug fixes.

DISTRIBUTIONS FOR THE RESPONSE

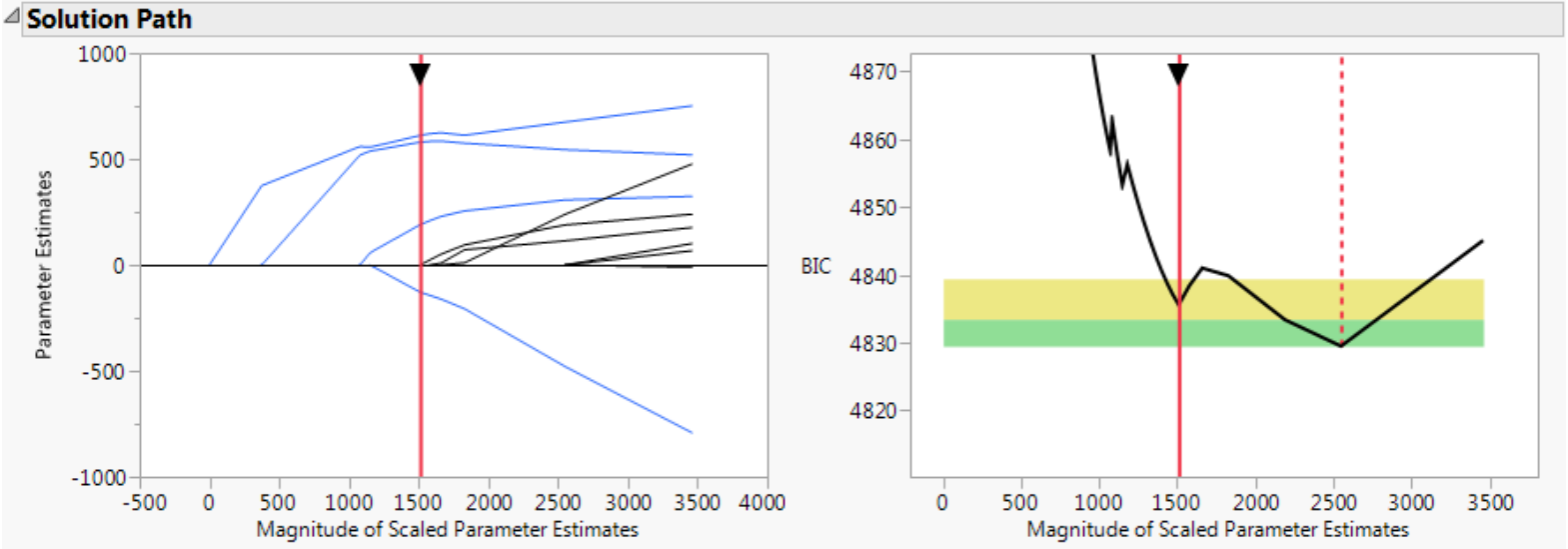
Many more distributions added in JMP 12:





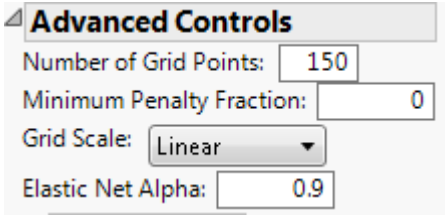
INTERACTIVE SOLUTION PATH

- Allows you to explore different models by changing the penalty setting.
- For KFold, Leave-one-out, BIC and AIC, you also get some guidance for equivalent models.



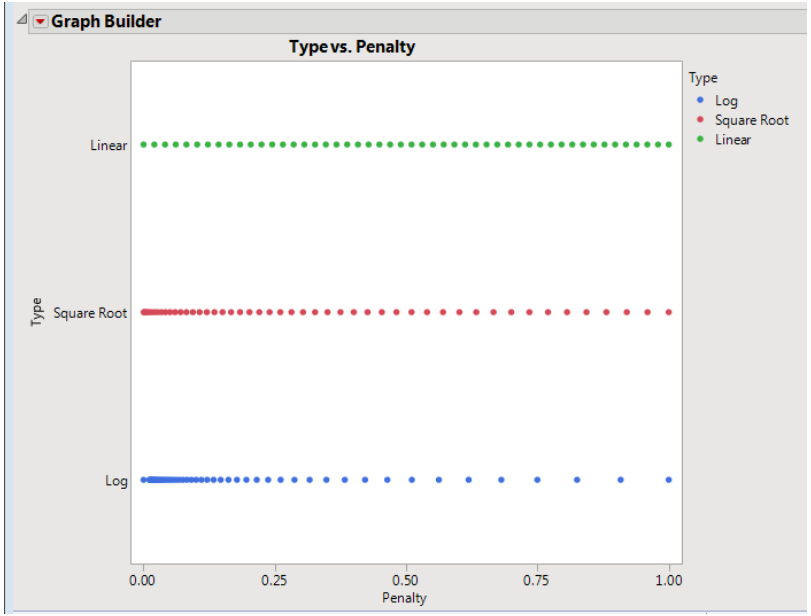
ADVANCED CONTROLS

- Advanced controls allow setting:
  - Number of grid points
  - Scale of the grid points
  - Minimum penalty fraction (*p*)
  - Elastic Net alpha (balance between l_1 and l_2 penalties)



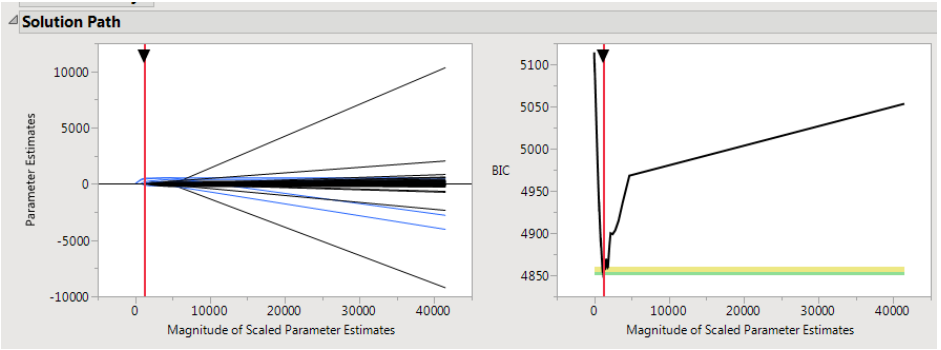
ADVANCED GRID CONTROLS

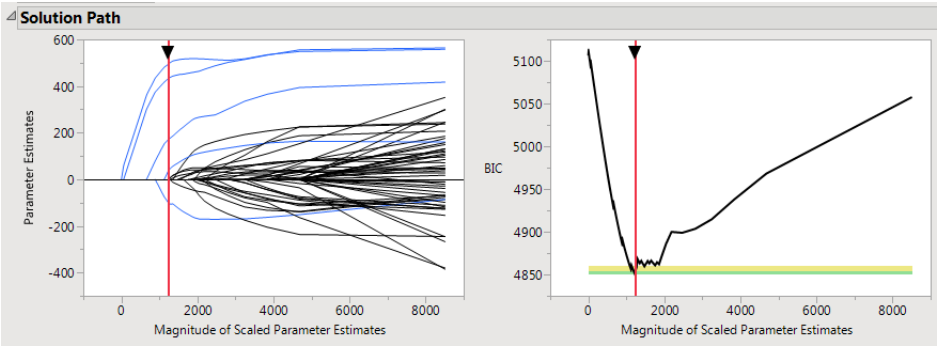
- The fitting routine searches over a grid of λ values to find the optimal penalty.
- JMP12 allows user to control:
  - Number of grid points
    - Default is 150 points
  - Scale of the grid points
    - Linear, Log, or Square Root scale
    - Default is Linear



OTHER ADVANCED CONTROLS

- Minimum penalty fraction (p)
  - $p = 0 \Rightarrow$ no penalty $\Rightarrow$ MLE
  - If MLE is overfitting, solution path can be distorted!
- Elastic Net alpha
  - Balances between l_1 and l_2 penalties
  - Avoids searching over a 2nd grid
  - $\alpha = 0 \Rightarrow$ Ridge (all l_2 penalty)
  - $\alpha = 1 \Rightarrow$ Lasso (all l_1 penalty)
  - Default=0.9 (variable selection with some singularity handling)





EXAMPLES

- Heart Disease Data
- Bowl Game Data

CONCLUSIONS

- The Generalized Regression platform allows users to build models that are easy to interpret and that predict well.
  - Even when the response isn't normally distributed
  - … and when the predictors are correlated
  - … and even when we don't have enough observations!

- Many new features and improved computational time will be included in JMP Pro version 12.

REFERENCES

- Burnham, K. and Anderson, D. 2002. *Model Selection and Multimodel Inference*. New York, NY: Springer.
- Gregg, X. “Best teams for college bowl attendance & TV ratings,” http://blogs.sas.com/content/jmp/2013/12/06/most-and-least-desirable-bowl-teams/ (accessed 15Aug2014).
- Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning*, Second Edition, New York, NY: Springer.
- Hesterburg, T., Choi, N., Meier, L., and Fraley, C. 2008. “Least angle and l_1 penalized regression: A review,” *Statistical Surveys*, 2, p. 61–93.
- Hoerl, A. and Kennard, R. 1970. “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” *Technometrics*, 12:1, p. 55–67.
- McCullagh, P. and Nelder, J. 1989. *Generalized Linear Models*, Second Edition, London: Chapman & Hall.
- SAS Institute Inc. 2015. *JMP® 12 Fitting Linear Models*. Cary, NC: SAS Institute Inc.
- Tibshirani, R. “Model selection and validation 2: Model assessment, more cross-validation,” http://www.stat.cmu.edu/~ryantibs/datamining/lectures/19-val2.pdf (accessed 15Aug2014).
- Tibshirani, R. 1996. “Regression Shrinkage and Selection via the Lasso,” *JRSSB*, 58:1, p. 267–288.
- von Bortkiewicz, L. 1898. *Das Gesetz der Kleinen Zahlen*. Leipzig: Teubner.
- Zou, H. and Hastie, T. 2005. “Regularization and variable selection via the elastic net,” *JRSSB*, 67:2, p. 301–320.

THANK YOU!

Michael Crotty & Clay Barker
Research Statisticians
JMP Division, SAS Institute