Paper RIV-151

# Penalizing Your Models: An Overview of the Generalized Regression Platform

Michael Crotty, SAS Institute; Clay Barker, SAS Institute

## ABSTRACT

We provide an overview of the Generalized Regression personality of the Fit Model platform added in JMP Pro 11. The motivation for using penalized regression is discussed, and multiple examples show how the platform can be used for variable selection on continuous or count data. There is also a sneak peek of Generalized Regression features coming in JMP Pro 12.

Keywords: JMP Pro, generalized linear models, penalized regression, lasso, elastic net.

## INTRODUCTION

Generalized Regression was added as a new personality of the Fit Model platform in JMP Pro 11. This paper will provide an introduction to the capabilities of this new feature, motivation for its use, and a look at what additional features will be available for Generalized Regression when JMP Pro 12 is released in 2015. (While technically not a platform in JMP, we will refer to the Generalized Regression personality of the Fit Model platform as a platform. In a practical sense, the personalities of the Fit Model platform are all substantial enough to be considered platforms that just have a common launch dialog.)

The paper will follow this outline. The next section will provide motivation for why we need a more general form of linear regression. The next two sections will provide background information on generalized linear models and penalized regression, respectively. The next section details new features that have been added in the JMP Pro 11 maintenance releases and upcoming in JMP Pro 12. The next section contains selected results from examples demoed in the presentation of this paper. The final section is a brief conclusion.

## MOTIVATION

The first question the reader might ask is "Why do I need a new generalized form of regression?" Before answering that, we first start with a very brief description of linear regression and some of its aims. Linear regression is useful when one is presented with data containing a response and one or more variables that might be able to explain variation in the response. The basic regression model is given by:

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

There are many reasons that people build regression models. Here are two primary reasons:

1. To better understand data and make decisions.
2. To make predictions for new observations.

Fortunately, the Generalized Regression platform in JMP Pro can help users with both of these goals! It is useful both for making decisions based on your data as well as making predictions for new observations. There are really two major concepts involved in the Generalized Regression platform. It is fitting *penalized generalized* linear models. The models are *penalized* in the sense that we add a penalty to the likelihood of the model. Depending on the form of the penalty, this penalty allows us to do variable selection as well as shrinkage of estimators. The models are *generalized* in the sense that they do not require the assumption that the response is normally distributed. This is useful in practice, since there are many situations that give rise to responses that are not normally distributed: count data, binary (yes/no) data, skewness and outliers, to name a few.

## GENERALIZED LINEAR MODELS

As mentioned in the motivation section, even though a lot of regression tools assume the response variable is approximately normally distributed, there are a lot of situations where that assumption is not appropriate. Here are

some examples where the normality assumption is not appropriate: insurance claims are often skewed (leading to gamma distributed responses), presence of heart disease is yes/no (leading to a binomial distribution), and the number of horse kick victims in the Prussian army (leading to a Poisson distribution) [10]. Generalized linear models allow for the assumption of normally distributed responses to be relaxed. In JMP Pro 11, the Generalized Regression platform supports seven distributions for the response; in JMP Pro 12, the platform supports 14 distributions for the response. Table 1 lists the distributions available in the Generalized Regression platform. They are listed in three categories based on the type of response distribution: continuous, discrete and zero-inflated. The zero-inflated distributions are discrete distributions, with the exception of the gamma which is continuous. The *italicized* distributions are ones that are being added in JMP Pro 12.

Table 1: List of distributions available in the Generalized Regression platform, categorized by general type of response distribution. *Italicized* distributions are ones that are being added in JMP Pro 12.

| **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* |

Entire graduate level courses and textbooks exist to cover generalized linear models (GLMs), so we will only be making a brief introduction to them here. There should be enough here so that the user has a basic understanding of the models that the platform is fitting. The basic linear model is $E(Y) = \mu$ where $\mu = X\beta$. For GLMs, there is a linear predictor

$$\eta = \sum_{j=1}^{p} x_j \beta_j$$

that is related to the mean component through a link function $g$:

$$\eta = g(\mu).$$

For simple linear regression, the link function is the identity function: $g(\mu) = \mu$.

For further information on generalized linear models, we refer you to McCullagh & Nelder (1989) [6].

## PENALIZATION METHODS

Most of the time when you hear about a penalty, it's a bad thing. But, in the case of penalization methods in regression, a penalty can be a very good thing.

The typical way to fit many statistical models is using an estimation technique called maximum likelihood. Maximum likelihood gives us the model that best fits (is most likely) given the data that has been observed. If we optimize a penalized likelihood instead, we get certain benefits:

- We will predict better on new data (by avoiding overfitting).
- We may have a model that is easier to interpret.
- We can overcome certain problems with our data:

  Not enough observations (more columns than rows, the "$n < p$" problem)

  Correlated predictors (multicollinearity)

To maximize the likelihood function $L(\beta)$, we usually reformulate the maximization problem to be a minimization problem where we seek to minimize the objective function $Q(\beta) = -\log L(\beta)$. To penalize the likelihood, we simply

add a penalty term to this objective function to create a penalized objective function:

$$Q(\beta) = -\log L(\beta) + \lambda \sum_{j=1}^{r} p(\beta_j)$$

where $\lambda$ controls the stiffness of the penalty and the form of $p(\beta_j)$ determines the flavor of penalized regression that is performed.

Note that when $\lambda = 0$, the penalty term goes away and we are left with the maximum likelihood objective function. As $\lambda$ increases, the effect of the penalty term on the objective function also increases.

Many penalties have been proposed in the literature, but there are three available in JMP Pro. They are the Lasso [9], Ridge [5] and Elastic Net [11]; the form of the penalty terms for these three methods are shown in Table 2.

Table 2: List of penalization methods and associated penalties available in JMP Pro.

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

The key idea for understanding why we might want to use penalized estimates in regression is that we are willing to accept some bias in order to reduce variance. For example, imagine two estimators for the probability of a coin landing on heads. The first is an estimator with uniform probability between zero and one; this will be unbiased, but the variance is larger than you might like. The second is an estimator normally distributed with a mean of 0.55, but with a standard deviation of less than 0.1; this estimator is biased, but the variance is much smaller than the variance of the first estimator.

Another problem in many regression problems that penalized regression can help us with is having lots of variables. In standard linear regression, this situation can easily cause you to overfit. This means that the model will fit the observed data very well, but it will perform poorly on new observations. There are two ways that penalization methods can help here: selection and shrinkage. The Lasso and Elastic Net shrink some predictors all the way to zero; this provides us with variable selection. Both selection and shrinkage help us avoid overfitting. Having lots of variables can also lead to the $n < p$ situation where the maximum likelihood method cannot be fit; in this case, penalized likelihood methods are still applicable. This situation arises when you have "short and wide" data.

At this point, you might be convinced that penalizing your regression model is a good idea but wondering how to pick between the various methods. Table 3 seeks to provide some guidance as to how to choose a penalization method. Maximum likelihood is the unpenalized fit and therefore provides no variable selection or shrinkage. Forward selection provides variable selection but no shrinkage. Ridge regression shrinks the estimators but not all the way to zero, so there is not variable selection in ridge regression. Lasso and Elastic Net both provide both shrinkage and variable selection. To help choose between the two methods, there are two considerations to take into account. The Lasso will tend to give you a more parsimonious model than the Elastic Net, while the Elastic Net can better handle collinearity than the Lasso.

Table 3: Relationship between the penalization methods available in JMP Pro with regards to variable selection and shrinkage.

| | | **Selection** | |
|---|---|---|---|
| | | No | Yes |
| **Shrinkage** | No | ML | Forward Selection |
| | Yes | Ridge | Lasso & Elastic Net |

In the Generalized Regression platform, the key to understanding the model (and the model fitting process) is the solution path graph. Figure 1 shows an example of a solution path graph with the nonzero terms highlighted. The red line shows the model (value of the penalty) that was chosen based on the chosen validation method. Each line in the plot corresponds to a parameter ($\beta_j$) in the model. You can think of the penalty being relaxed as you move along the horizontal axis from left to right. In most cases, the maximum likelihood estimate (penalty equal to zero)

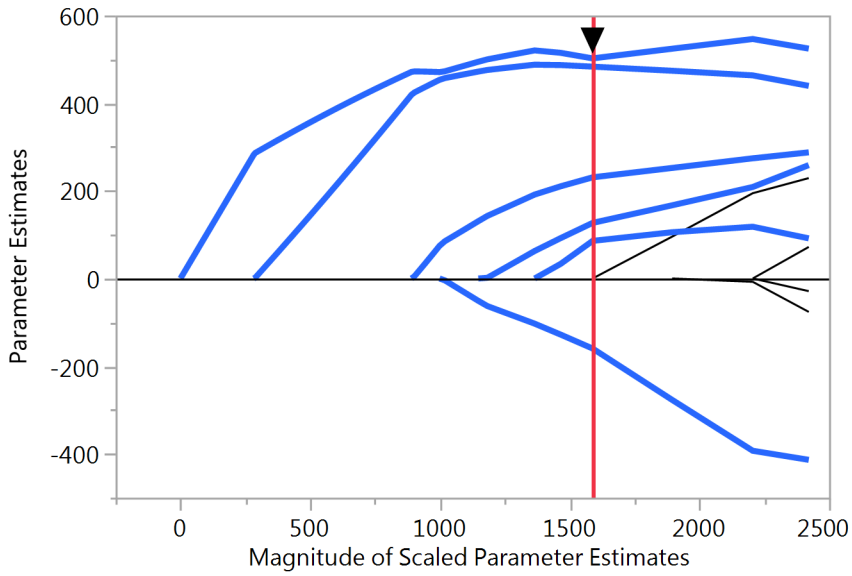



Figure 1: Example solution path graph in JMP Pro with nonzero terms highlighted.

is at the far right of the plot. The points of the lines are plotted at the grid points used to search for the optimal penalty value; the vertical position of those points is determined by the value of the parameter estimate for each line at each grid point of the penalty value. In JMP Pro 12, the red line is draggable so that you can explore other models based on different values of the penalty term. As you drag the red line, the report will update to show the results for different penalty values.

The Lasso and Elastic Net both have an "adaptive" version which use a modified version of the $l_1$ penalty. This modified version uses weights generated from the original maximum likelihood estimates. This allows us to use some information about which predictors are strong predictors of the response to improve our model fit.

The final topic to cover in this section is the platform's use of validation methods. The Generalized Regression platform has many different methods for validation. These are used to determine the optimal value of the penalty in penalized regression. We will not go into detailed explanations here, but we refer you to the JMP documentation [7] for more information on these methods. Here is a list of the methods currently available:

- Kfold
- Holdback
- Leave-One-Out
- BIC (not available for Ridge)
- AIC (not available for Ridge)
- None (only available for Maximum Likelihood)
- Validation Column

The BIC method is the default because it is usually the fastest and it generally results in a parsimonious model. Note that maximum likelihood can only use the last two methods, because there is no penalty to optimize in maximum likelihood.

For further information on penalized regression models, we refer you to a very good overview paper of penalized regression by Hesterberg et al [4].

## NEW FEATURES

This section briefly describes the new features and enhancements that have gone into the Generalized Regression platform in the maintenance releases of JMP Pro 11 and the features and enhancements planned for JMP Pro 12 (to be released in 2015).

### JMP PRO 11.1

The main focus of work on the Generalized Regression platform for JMP Pro 11.1 was improving computation time. In 11.1, the computation time was sometimes improved dramatically.

The lines in the solution path plot became selectable. This is useful for identifying which term a line in the solution path refers to. If a line is a long way from the others in the path, it is easily identified by clicking on it in the plot. Clicking on it not only selects the line in the plot; it also selects the row in the parameter estimates table in the Generalized Regression report window for the term corresponding to the selected line in the plot. This selection also selects the corresponding column in the data table.

Early stopping rules for the Lasso, Ridge and Elastic Net estimation methods were added in 11.1 as well. These rules provide another way to speed up the fitting process, especially for large problems. As the algorithm moves along the grid of penalty values, early stopping will cause the algorithm to terminate when it has hit ten consecutive grid points where the fit has not improved upon the best fit as determined by the validation method. The method for Elastic Net is slightly more complex since there are two penalty terms over which to optimize.

As with any release of a new version of software, there were a variety of bugs fixed in the Generalized Regression platform for JMP Pro 11.1.

### JMP PRO 11.2

Further work for improving computation time occurred in JMP Pro 11.2.

The drawing of the solution path was improved in that the $y = 0$ line is no longer selectable; this helps with the interactive selection of lines in the solution path plot (added in 11.1).

The platform also added support for three-level validation columns, corresponding to having training, validation and test sets.

As with any release of a new version of software, there were a variety of bugs fixed in the Generalized Regression platform for JMP Pro 11.2.

### JMP PRO 12

Another big boost to the speed of the platform is expected in JMP Pro 12 (to be released in 2015) by utilizing a new faster numerical optimizer. Seven new response distributions are being added as denoted in Table 1.

A second interactive solution plot with validation information across the grid points of the penalty is being added. For most validation methods, this plot will assist the user in choosing an appropriate model when many models around the optimal model are statistically equivalent. Figure 2 shows an example of such a plot. With the BIC and AIC validation methods, there are two regions of guidance based on rules of thumb provided by Burnham & Anderson (2002) [1]; the two regions provide guidance to help determine which models are close to the optimal model, with two levels of support. In Figure 2, the model selection line has been moved to the left (fewer terms in the model) to a model with a BIC in the yellow shaded region; this can provide a more parsimonious model that is still determined by BIC validation to be a relatively good model. For the KFold and Leave-one-out validation methods, the plot has a single shaded region for the scaled negative likelihood within which models are statistically equivalent, based on the one-SE rule [8].

Many advanced controls for the Lasso and Elastic Net estimation methods are being added. In JMP Pro 12, the user will be able to set the number of grid points used in the search for the optimal penalty value; by default, the platform uses 150 grid points. Prior to JMP Pro 12, the grid points were always spaced equally (linearly), but now the user will be able to choose the scale of the grid points (between Linear, Log and Square Root). By default, when it is possible, the penalty will go all the way to zero (the maximum likelihood fit). Now, there is an option to set the minimum penalty fraction, which will allow the user stop the fitting process prior to reaching the maximum likelihood fit. This will be useful when the data contains a singularity and in cases where the vertical scale of the solution path gets distorted because the paths go all the way to the maximum likelihood estimate. Finally, the Elastic Net has a new advanced control setting that allows the user to select the Elastic Net $\alpha$ level; the $\alpha$ is the parameter that

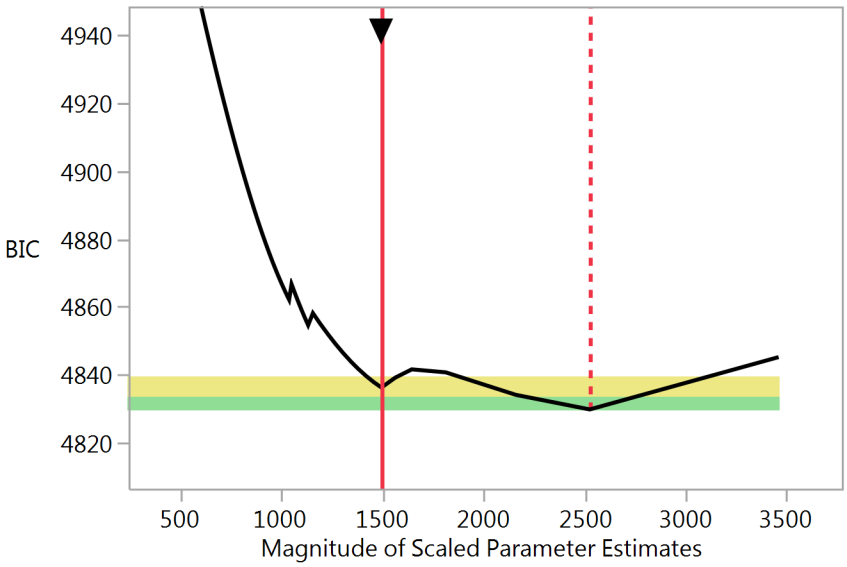



Figure 2: Example BIC path graph in JMP Pro 12 with a more parsimonious model selected.

determines the balance between the $l_1$ and $l_2$ penalties. The default setting for the Elastic Net $\alpha$ is 0.9; this setting provides for selection with some singularity handling.

Many other new features are being added in JMP Pro 12 as well. There will be new diagnostic plots for exploring residuals. Inverse prediction and multiple comparisons will be supported similar to other Fit Model personalities. The user will have the ability to force specific terms into the model, and the platform will support forward selection.

As with any release of a new version of software, there are a variety of bugs being fixed in the Generalized Regression platform for JMP Pro 12.

## EXAMPLES

This section describes two examples presented in the live presentation of this paper at SESUG 2014. Results can be obtained using the JMP journal file associated with this paper, available from the authors upon request.

### HEART DISEASE DATA

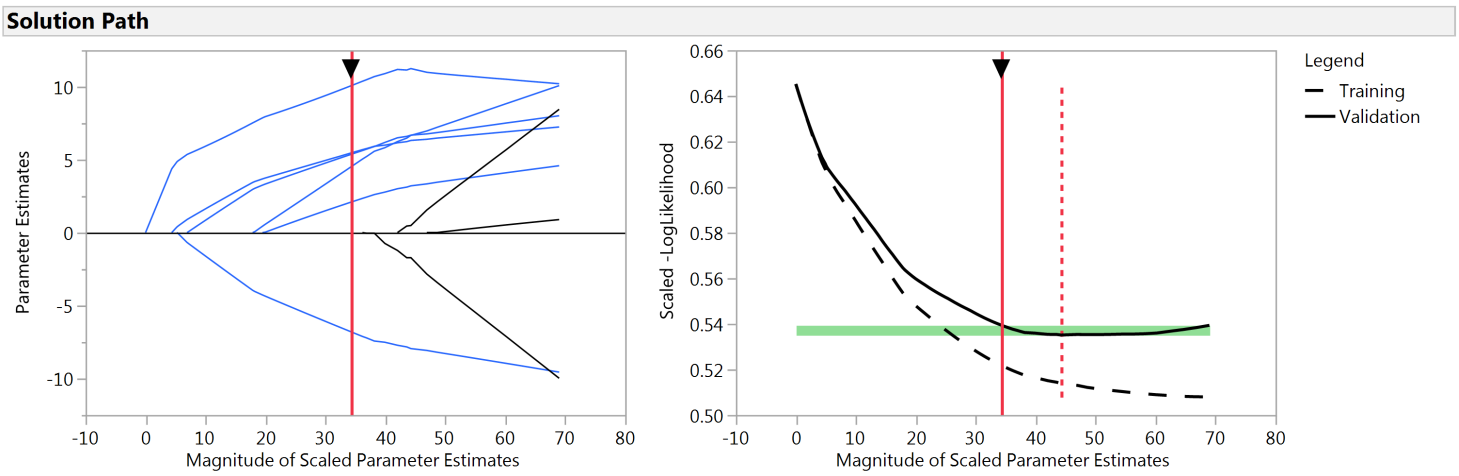
Our first example uses heart disease data from a study of South African males, recounted in Hastie et al (2009) [3]. The response in this data is a binary response indicating if a patient has coronary heart disease or not. Therefore, using a normal distributed response model would not be appropriate. The predictor variables include a variety of measures of patient health, family history of heart disease, behavior, obesity, alcohol consumption and age. We are interested in using the Generalized Regression platform to fit a binary response model to a parsimonious set of variables for these data.

To do this, we use the Generalized Regression platform with a Binomial response distribution. In the Model Launch dialog, we choose a non-adaptive Lasso penalty and KFold validation (with 5 folds). Figure 3 shows the results of the analysis. We elect to move the solution path from the optimal solution to a more parsimonious one that is still in the shaded region based on the one-SE rule.

We can see from the results here that Adiposity, Obesity and Alcohol Consumption have dropped out of the model and do not contribute to predictions for future observations. From this point, the remaining effects in the model can be explored using the Prediction Profiler available in most JMP modeling platforms, including Generalized Regression.

### BOWL GAME DATA

Recently in the JMP development group, a challenge [2] was posed to determine desirability scores for American college football teams in postseason bowl games. Desirability is determined by bowl game attendance and/or




**Parameter Estimates for Centered and Scaled Predictors**

| Term | Estimate | Std Error | Wald ChiSquare | Prob > ChiSquare | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | -0.740546 | 0.1143924 | 41.909275 | <.0001* | -0.964751 | -0.516341 |
| Systolic Blood Pressure | 2.0926493 | 2.5139892 | 0.6928929 | 0.4052 | -2.834679 | 7.0199776 |
| Tobacco | 5.4609482 | 2.4452093 | 4.9877427 | 0.0255* | 0.6684261 | 10.25347 |
| LDL Cholesterol | 5.3658395 | 2.3123645 | 5.3847157 | 0.0203* | 0.8336884 | 9.8979907 |
| Adiposity | 0 | 0 | 0 | 1.0000 | 0 | 0 |
| Family History[Absent] | -6.808179 | 2.4111098 | 7.9731133 | 0.0047* | -11.53387 | -2.08249 |
| Type A personality | 4.5190079 | 2.4333373 | 3.4489076 | 0.0633 | -0.250246 | 9.2882614 |
| Obesity | 0 | 0 | 0 | 1.0000 | 0 | 0 |
| Alcohol Consumption | 0 | 0 | 0 | 1.0000 | 0 | 0 |
| Age at onset | 10.068512 | 3.0140522 | 11.159098 | 0.0008* | 4.1610787 | 15.975946 |

Figure 3: Results of Binomial Lasso fit for the South African Heart Disease data.

television ratings. One method for approaching this challenge using Generalized Regression is described here.

The data initially presented for the challenge contained factors such as Year, Bowl, Teams, Location, television rating, attendance and location data for the bowl game and participating teams. We augment the data to create indicator columns for each team; hence, each row (corresponding to a bowl game in a particular year) has zeros in the indicator columns with the exception of the two columns corresponding to the two teams playing that bowl game in that year. Once the data are in this form, we use Generalized Regression to do variable selection on these indicator columns to find their relative importance to a model predicting a desirability measure composed of a weighted average of attendance and television ratings.

We also include the bowl game as a factor, since there is a lot of variation in attendance and television ratings among the various bowl games.

To perform the analysis, we do two separate Lasso fits and then combine the results, taking care to normalize the parameter estimates from each fit so that the attendance and television ratings ranking values are on the same scale.

The resulting analysis produces a small number of nonzero team parameters that we conclude are the most influential teams with regards to attendance and television ratings. These nonzero parameters are mostly positive, but there are some negative parameters, suggesting that those teams negatively affect attendance and/or television ratings. The bowl game effect is also significant in the final model, as we would expect.

## CONCLUSION

We hope this paper has helped the reader understand some of the background of the Generalized Regression platform in JMP Pro and why it might be useful. Background information and motivation was provided. Descriptions of the enhancements added in the JMP Pro 11 maintenance releases and coming soon in JMP Pro 12 (to be released in 2015). We also include the disclaimer that all features mentioned for JMP Pro 12 are still under development at the time of writing this document, and as such, they are subject to change prior to the final release of JMP Pro 12.

We concluded with two examples of using the Generalized Regression platform.

## REFERENCES


[1] Burnham, K. and Anderson, D. 2002. *Model Selection and Multimodel Inference*. New York, NY: Springer.

[2] 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).

[3] Hastie, T., Tibshirani, R., and Friedman, J. 2009. "The Elements of Statistical Learning", Second Edition, New York, NY: Springer.

[4] 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.

[5] Hoerl, A. and Kennard, R. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12:1, p. 55–67.

[6] McCullagh, P. and Nelder, J. 1989. "Generalized Linear Models," Second Edition, London: Chapman & Hall.

[7] SAS Institute Inc. 2015. "JMP® 12 Fitting Linear Models." Cary, NC: SAS Institute Inc.

[8] 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).

[9] Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso," *JRSSB*, 58:1, p. 267–288.

[10] von Bortkiewicz, L. 1898. *Das Gesetz der Kleinen Zahlen*. Leipzig: Teubner.

[11] Zou, H. and Hastie, T. 2005. "Regularization and variable selection via the elastic net," *JRSSB*, 67:2, p. 301–320.


## CONTACT INFORMATION

Your comments and questions are valued and encouraged.


Contact the author(s):

| | |
|---|---|
| Name | Michael Crotty |
| Enterprise | SAS Institute |
| Address | SAS Campus Drive |
| City, State, ZIP | Cary, NC 27513 |
| E-mail: | `mailto:michael.crotty@sas.com` |